

大数据应用分析技术与方法

刘汝焯 戴佳筑 何玉洁 编著

清华大学出版社

大数据应用分析技术与方法

刘汝焯　戴佳筑　何玉洁　编著

清 华 大 学 出 版 社

北　京

内 容 简 介

本书强调了大数据的宝贵价值，介绍了常用的数据分析技术与方法，论述了大数据分析的思维特征，紧扣大数据的特点演示了可视化分析与可视化挖掘的方法，详细讨论了数据清洗与元数据管理，对大数据的风险予以充分揭示，同时提出了大数据风险管理的对策，对大数据治理作了简介。

本书具有很强的实用性、可操作性和指导性，对于企业管理人员、企业数据分析人员、业务分析人员和市场营销人员，政府监管机构如证监会、银监局、保监会的监管人员，审计师、注册会计师，纪检监察和司法机关执纪执法人员有参考价值，同时可供高等院校相关专业的师生参阅。

图书在版编目（CIP）数据

大数据应用分析技术与方法/刘汝焯，戴佳筑，何玉洁编著. —北京：清华大学出版社，2018
ISBN 978-7-302-48707-4

Ⅰ. ①大… Ⅱ. ①刘… ②戴… ③何… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 265959 号

责任编辑： 王　青
封面设计： 何凤霞
责任校对： 宋玉莲
责任印制： 杨　艳

出版发行： 清华大学出版社
网　　址： http://www.tup.com.cn，http://www.wqbook.com
地　　址： 北京清华大学学研大厦 A 座　　**邮　　编：** 100084
社 总 机： 010-62770175　　**邮　　购：** 010-62786544
投稿与读者服务： 010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈： 010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者： 三河市铭诚印务有限公司
经　　销： 全国新华书店
开　　本： 185mm×260mm　　**印　　张：** 13　　**字　　数：** 299 千字
版　　次： 2018 年 1 月第 1 版　　**印　　次：** 2018 年 1 月第1 次印刷
印　　数： 1～3000
定　　价： 39.00 元

产品编号：065688-01

前　言

随着大数据的迅猛发展和日益普及，越来越多的与数据分析有关的人员，如企业管理人员、企业数据分析人员、业务分析人员、市场营销人员，政府监管机构如证监会、银监局、保监会的监管人员，审计师、注册会计师，纪检监察和司法机关执纪执法人员等需要掌握大数据应用分析技术与方法，迫切需要从大数据中挖掘有用的信息，提升工作水平和工作效率。这是信息化发展提出的必然要求，尤其是在业务与信息技术密切融合的形势下，这种需求越来越强劲。适应这种需求，我们在编著这本书时，着力突出实用性、可操作性和指导性。

实用性。贴近大数据发展的现状和趋势，全书共安排了10章内容，强调了大数据的宝贵价值，介绍了常用的数据分析技术与方法，紧扣大数据的特点演示了可视化分析与可视化挖掘的方法，详细讨论了数据清洗与元数据管理，对大数据的风险予以充分揭示，同时提出了大数据风险管理的对策，对大数据治理作了简介。

可操作性。在介绍大数据应用分析技术和方法时，由浅入深，逐步引导，屏蔽技术细节，让读者直接进入业务应用的层面，熟练掌握操作。尤其是全书从大数据分析的应用实践中精选了大量案例，进行了生动讲解。这些案例是大数据分析实践中的可贵探索和经验总结。通过案例的操作可以更好地引导读者加深对理论部分的理解，掌握分析技术与方法。

指导性。本书创新性地把大数据应用分析划分为器、技、道和美四个层面。器，指大数据分析的硬件和软件；技，指大数据分析的技术和方法；道，指大数据分析的思维方式；美，指审美体验、感觉和想象力。在开展大数据分析时需要硬件和软件、需要技术方法，这是毋庸讳言的。但大数据数量巨大、类型繁多、来源复杂，而且很多过去从来没有遇到过，单靠工具和技术方法是不能胜任大数据分析的多变情况的，清晰的分析思路、科学的思维方式显得更为重要，具有更强的更普遍的指导性。本书详细介绍了特征发现的思维方式，通过案例介绍了特征枚举、特征捕捉、特征分析的实际应用，同时对大数据分析中如何结合审美体验，张开想象的翅膀，激发分析的灵感，打开分析的思路给予了必要强调。

为天天和大数据打交道的人尽快掌握大数据分析的实用技能助一臂之力，为天天使用大数据的人通过最简单的路径掌握大数据分析的技能提供支持和帮助，这是我们的初衷。为了这个初衷我们确实努力了。大数据在发展中，加之编著者的水平和经验有限，书中有些问题的研究还不透彻，有些内容还有待于在实践中检验和完善，还有些可能本身就是存在问题的，这也是在所难免的。真诚希望广大读者批评指正！

参加本书撰写的有刘汝焯、戴佳筑、何玉洁。刘汝焯设计了全书的章节，撰写了其中第 2、4、7 章，对全书进行了统稿。戴佳筑撰写了第 1、3、6、10 章和附录 B。何玉洁撰写了第 5、8、9 章和附录 A，对全书的书稿进行了统一修订。

刘汝焯

2017 年 6 月 11 日于北京

目　　录

第1章　大数据是信息社会的宝贵资源

1.1　大数据产生的背景和概念

大数据是随着信息数据快速增长和网络计算技术迅猛发展而兴起的一个新概念。大数据通过对海量数据的收集、处理和展示，揭示规律，预测未来。大数据能够帮助企业从海量数据中挖掘用户的需求，从而使数据真正产生价值。随着大数据的发展，其应用已经渗透到农业、工业、商业、服务业和医疗领域等各个方面。

随着计算机信息技术的发展和网络的普及，以博客、社交网络、基于位置的服务为代表的新型信息发布方式的不断涌现，以及云计算、物联网、移动互联网等技术的兴起和普及，数据正以前所未有的速度在不断地增长和累积，特别是进入DT(数据技术)时代，在线数据存储和计算量以及人类在日常学习、生活、工作中产生的数据量正以指数形式增长，呈现“爆炸”状态。国际数据公司(IDC)的研究结果表明，2008年全球产生的数据量为0.49ZB(1024GB=1TB，1024TB=1PB，1024PB=1EB，1024EB=1ZB)，2009年的数据量为0.8ZB，2010年增长为1.2ZB，2011年的数量更是高达1.82ZB，相当于全球每人每年产生200GB以上的数据。而到2012年为止，人类生产的所有印刷材料的数据量是200PB。2014年，全球产生的数据量估计已经达到了3.6ZB。

全球信息数据量的飞速膨胀成为大数据产业存在并发展的基础。国际数据公司(IDC)预计，未来全球数据总量增长率将维持在50%左右，到2020年全球数据总量将达到40ZB，其中，我国将达到8.6ZB，占全球的21%。中国信息产业研究院的数据显示，2014年我国大数据市场规模约为116亿元，同比增长38%。预计未来几年，随着应用效果的逐步显现，我国大数据市场规模还将维持40%左右的高速增长。

除了迅速增长的数据洪流，数据的结构越来越趋于复杂化，除了传统数据库中的数据，还有文档、网页、图像、音频和视频等，而且后者所占的比例也越来越大。这些数据的量变到底有多大呢？2014年产生了大约5ZB(Zettabyte)字节的非结构化数据，到2020年预计将增加到大约40ZB字节的非结构化数据。如图1-1所示为非结构化数据2005—2020年的实际和预期增长对比，该图片引自Evangelos Simoudis的“认知应用：大数据的下个转折点”一文。

这些数量巨大、种类繁多、结构复杂的数据早已远远超越了传统技术所能处理的范畴，如何合理、高效、充分地管理和使用这些数据，使之能够给人们的生活和工作带来更大的效益和价值，逐渐成为人们的共识，在这种背景下，大数据应运而生。

什么是大数据呢？大数据一词源于英文的“Big Data”，以前也有类似的词语，如“海量数据”“信息爆炸”等，但似乎都很难准确描述这个词的具体内涵。目前国内外对大数据

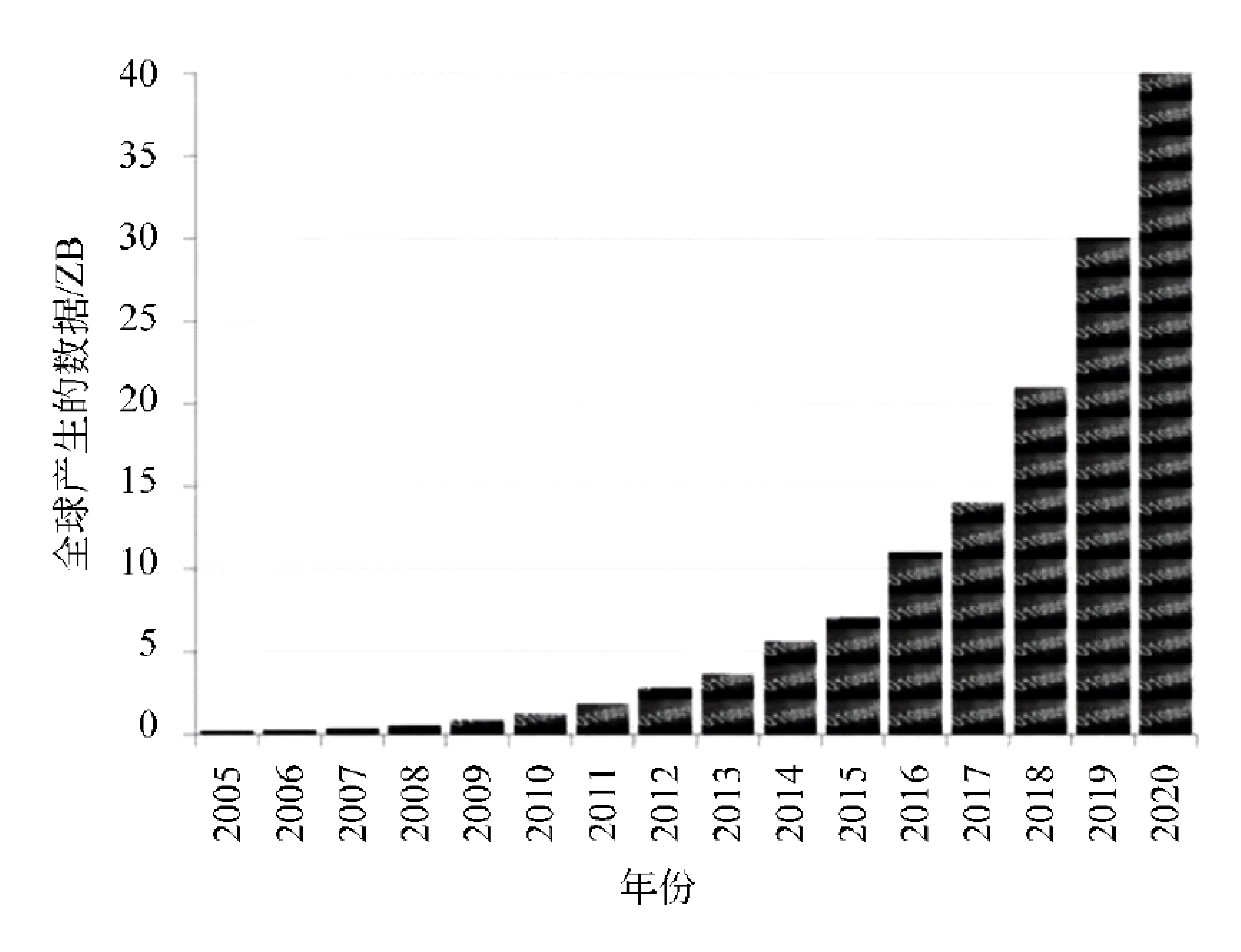


图 1-1　非结构化数据 2005—2020 年的实际和预期增长对比

没有一个统一的定义，国内外政府机构、企业和专家从不同角度给出了大数据的定义。维基百科对大数据的定义是“大数据是数据规模巨大，通过目前主流软件工具无法在合理时间内捕获、管理、处理并整理成为帮助经营决策的数据集”；美国国家标准和技术研究院(NIST)则认为“大数据是指由于数据的容量、数据的获取速度或者数据的表示限制了使用传统关系方法对数据的分析处理能力，需要使用扩展的机制以提高数据处理效率的技术”；著名的管理咨询公司麦肯锡公司的研究报告中将大数据定义为“超过了传统数据库软件工具捕获、存储、管理和分析能力的数据集”；国际数据公司(IDC)是研究大数据及其影响的先驱，在其 2011 年的报告中指出“大数据技术描述了一个技术和体系的新时代，被设计用于从大规模、多样化的数据中通过高速捕获、发现和分析技术提取数据的价值”。著名的大数据专家维克托·迈尔—舍恩伯格在其经典著作《大数据时代》中，指出大数据“是当今社会所独有的一种新型能力，以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见。”

大数据中的海量数据有三个主要来源，首先是海量交易数据。随着信息技术的广泛应用，越来越多的企业和机构比以往任何时候都依赖信息系统，如超市的销售记录系统、火车售票系统、银行的交易记录系统、医院病人的医疗记录等，由此产生了大量的交易数据。其次是海量的网络信息。互联网的诞生促使人类社会数据量出现一次巨大的飞跃，但是真正的数据爆发产生于移动互联网时代特别是社交媒体的兴起，这类数据近几年一直呈现爆炸性的增长，涵盖了海量的聊天记录、Web 网页、电子邮件、图片、视频、音频等。最后是海量的感知数据。物联网(The Internet of Things)是新一代信息技术的重要组成部分，通过传感器和网络技术实现了物与物、人与物、人与人之间的互联。物联网时代，除了智能手机、平板电脑等常见的客户终端之外，更多更先进的传感设备和智能设备，如智能手表、智能眼镜、智能汽车、智能电视、工业设备和手持设备等都将接入网络，由此产生的海量感知数据量及其增长速度比以往任何时期都要多。

近几年，大数据迅速成为科技界和企业界甚至世界各国政府关注的热点，发展的势头

不可阻挡。著名的科技旗舰杂志《自然》和《科学》等相继出版专刊，分别从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面专门探讨大数据带来的机遇和挑战。2011 年 5 月麦肯锡公司在美国拉斯维加斯举办的第 11 届 EMC World 年度大会上称："数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于大数据的挖掘和运用，预示着新一波生产力增长和消费盈余浪潮的到来。"美国政府认为大数据是"未来的新石油"，并于 2012 年 3 月 29 日发布了"大数据研究发展倡议"，正式启动"大数据发展计划"。

我国十分重视大数据的发展。2012 年 8 月，中国科学院启动了"面向感知中国的新一代信息技术研究"战略性先导科技专项，2013 年，科技部正式启动 863 项目"面向大数据的先进存储结构及关键技术"，这些科研项目的任务之一就是研制用于大数据采集、存储、处理、分析和挖掘的未来数据系统。国务院于 2014 年 8 月发布了《国务院关于加快发展大数据、呼叫中心等生产性服务业的指导意见》，从国家层面推动大数据的建设和发展；2015 年 7 月，国务院办公厅印发了《关于运用大数据加强对市场主体服务和监管的若干意见》，要求在政府层面推动大数据的应用；2015 年 9 月国务院颁布了《国务院关于印发促进大数据发展行动纲要》，提出"全面推进我国大数据发展和应用，加快建设数据强国"的方针政策，这标志着大数据建设和应用已经上升为国家发展的长期战略。

1.2　大数据的特征

目前大数据尚未具有统一的描述，不同的定义基本上都是从大数据的特征出发，通过大数据特征的阐述和归纳试图给出其定义。大家都普遍认同大数据具有 4 个基本特征，分别是容量大（Volume）、种类多（Variety）、高速度（Velocity）和价值密度低（Value），由于这四个特征的英文单词都是以英文字母"V"开头，所以又将其称为大数据的"4V 特征"。

容量大是指大数据的数据量非常巨大。例如，互联网搜索的巨头谷歌现在能够处理的网页数量是在千亿以上，每月处理的数据量超过 400PB（400×10^{15} B），并且呈继续高速增长的趋势；百度目前数据总量接近 1 000PB（$1\ 000\times10^{15}$ B），存储网页数量接近 1 万亿，每天大约要处理 60 亿次搜索请求。

种类多是指大数据的数据种类繁多，结构复杂。在大数据时代，数据来源并非仅仅是计算机产生的信息或者人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递有关位置、运动、振动、温度、湿度乃至空气中化学物质的变化等，也产生了海量的数据信息。这些数据既包含传统关系数据库中保存的结构化数据，也包含图像、声音和视频等非结构化数据以及 HTML 网页和 XML 文档等半结构化数据，而且非结构化数据和半结构化数据所占的比例呈现越来越大的趋势。

高速度是指大数据能够更快地满足实时性的需求。目前，对于数据智能化和实时性的要求越来越高，比如开车时会随时通过智能导航仪查询最佳路线，在餐厅吃饭时会查询其他用户对餐厅的评价和推荐的菜肴，见到有趣的事情或可口的食物会拍照发微博等诸如此类的人与人、人与机器之间的信息交流互动，这些都不可避免地带来数据交换，而数

据交换的关键是降低延迟，以近乎实时的方式完成数据交换的任务。

价值密度低是大数据特征里最关键的一点。数据量大并不意味着数据价值的增加，大数据时代数据的价值就像沙里淘金，其应用价值（金子）是隐藏在沙子之中的，数据量越大，里面真正有价值的东西所占的比例就会越少。大数据面临的一个挑战就是从这些TB、PB、EB级的海量数据中，提取有价值的信息，将信息转化为知识，发现规律，最终用知识促成正确的决策和行动。

另外，随着人们对大数据的研究不断深入，有的企业（如IBM公司）认为大数据还应具有第五个特征，即真实性（Veracity），通俗地讲，它是指大数据中数据来源广泛、种类繁多，这些数据具有不可靠或不精确的可能性。当我们试图获得大规模的数据时，必须能够控制这些不可靠或不精确带来的影响，使这些海量数据能够被用来更好地解释和预测客观世界。

1.3 大数据与传统数据的区别

从传统的数据库到大数据，不仅仅只是一个简单的技术演进，两者既有密切联系又有着本质上的差别。

大数据的出现颠覆了传统的数据管理方式，在数据来源、数据处理方式和数据思维等方面带来革命性的变化。为了说明传统的数据库和大数据的区别，有的专家使用“池塘捕鱼”和“大海捕鱼”的形象比喻。“鱼”是待处理的数据，“池塘捕鱼”代表传统数据库时代的数据管理方式，而“大海捕鱼”则对应着大数据时代的数据管理方式。“捕鱼”环境条件的变化导致了“捕鱼”方式的根本性差异，这些差异主要体现在如下几个方面。

（1）数据规模：“池塘”和“大海”最明显的区别就是规模不一样。“池塘”规模相对较小，“池塘”的处理对象通常以MB为基本单位，而“大海”的规模非常大，则常常以GB，甚至是TB、PB、EB为基本处理单位。

（2）数据类型：“池塘”中的数据种类往往仅仅有几种，这些数据又以结构化数据为主。而在“大海”中数据的种类繁多，这些数据不仅包含结构化数据，还包含半结构化数据以及非结构化的数据，并且半结构化数据和非结构化数据所占份额越来越大。

（3）模式和数据的关系：传统的关系数据库都是先有模式，然后才会产生数据。这就好比是先选好合适的“池塘”，然后才会向其中投放适合在该“池塘”环境生长的“鱼”。而大数据时代很多情况下难以预先确定模式，模式只有在数据出现之后才能确定，且模式随着数据量的增长处于不断的演变之中。这就好比“大海”中鱼的种类和数量都在不断地增长，鱼的变化会使大海的成分和环境处于不断变化之中。

（4）处理对象：在“池塘”中捕鱼，“鱼”仅仅是其捕捞对象。而在“大海”中，“鱼”除了是捕捞对象之外，还可以通过某些“鱼”的存在来判断其他种类的“鱼”是否存在。也就是说，传统数据库中数据仅作为处理对象，而在大数据时代，要将数据作为一种资源来辅助解决其他诸多领域的问题。

（5）处理方法：如果把“渔网”比作数据处理方法的话，捕捞“池塘”中的“鱼”，只需少数几种基本的“渔网”就可以应对，但是在“大海”中，不可能存在少数渔网能够捕获所有的

鱼类。传统意义上的数据处理方式包括数据挖掘、数据仓库、联机分析处理(OLAP)等,而在大数据时代,数据已经不仅仅是需要分析处理的内容,更重要的是人们需要借助专用的思想和手段从大量看似杂乱、繁复的数据中,收集、整理和分析数据,为人们在生产和生活中预测、决策和规划提供强有力的支持。

图灵奖获得者、著名数据库专家吉姆・格雷(Jim Gray)博士观察并总结在人类的科学研究史上,先后经历了实验、理论和计算三种研究方法。而在数据量不断增加和数据结构愈加复杂的今天,这三种方法在一些新的研究领域已经无法很好地发挥作用,所以吉姆・格雷博士提出了科学研究的第四种方法,即"数据探索",通过大数据的分析和处理来指导科学研究。

(6) 存储方式:"池塘"大都采用关系型数据库保存数据,而"大海"的数据量巨大,关系型数据库已经不能容纳如此巨大的数据,目前只能采用非关系型数据库(如 NoSQL)或分布式文件系统(HDFS)来存储数据。

虽然大数据和传统数据库有本质的差异,但是二者又有密切的联系。首先,大数据不是否定传统的数据库,有些学者认为传统数据库是大数据的一个重要组成部分,大数据只是传统数据库处理能力的拓展和延伸;其次,有些著名的 IT 企业提出传统数据库和大数据是互补的关系,大数据中的结构化数据通过传统数据库能够获得更好的存储和处理;最后,虽然传统的数据库在处理当今海量复杂的数据方面遇到了严峻的挑战,但是它依然是今天主流的数据存储技术,大数据要代替传统数据库成为主流的存储技术尚需时日。

1.4　大数据的价值和开发应用

近几年,大数据迅速发展成为政府、企业界和学术界关注的热点。人们意识到,一个国家和企业拥有数据的规模和运用数据的能力将成为综合国力和企业竞争力的重要组成部分,对数据的占有和控制将成为国家间和企业间新的争夺焦点。世界 500 强的大公司认为大数据是"重要的生产因素",而美国政府甚至把大数据称为"未来的新石油"。

毋庸置疑,大数据是待挖掘的金矿,其价值不言而喻。大数据的核心价值是什么呢?目前人们比较认同的有三个方面的价值。

首先,大数据改变了我们分析和使用数据的思维方式。《大数据时代》一书作者维克托・迈尔—舍恩伯格认为大数据时代处理和分析数据的思维有三大转变:第一个转变是在大数据时代可以分析更多的数据,甚至是相关的所有数据,而不再依赖少量的采样数据。在传统数据分析中,我们所做的是试图通过最少量的样本数据观测来发现规律。由于数据的采集、存储和分析的成本高,因此我们只能采用采样的方法。而在大数据时代,我们收集所有的数据,是与我们所研究的现象相关的所有可获得的数据,因此我们能够基于与某事物相关的所有数据展开数据分析,而不是仅仅依靠分析少量的数据样本。第二个转变是不再追求精确度。大数据时代数据是如此之多,以至于我们不再热衷于追求精确度。适当忽略数据的精确度,可以获得更广泛的数据,将带来更好的洞察力和更大的商业利益。第三个转变是不再热衷于寻找事物之间的因果关系,而是关注事物之间的相关关系。例如,成千上万的电子商务网站可以根据所记录的用户行为习惯,分析出用户喜爱

的产品或服务，然后对用户进行推荐，但是这些网站并不关心用户为什么会对这些产品和服务感兴趣。

其次，大数据提高了决策支持的能力。基于大数据的决策有两个主要特点：第一，不同于传统的基于少量数据样本的数据分析方法。大数据中的海量数据全面覆盖了企业经营以及政治、经济、社会、教育等方面的信息，通过对这些完整的信息进行分析，能够提高决策的质量；第二，决策的技术水平和效率大幅提高。云计算技术是大数据的重要支撑技术，通过云计算强大的计算能力和数据挖掘技术，人类不会被海量数据所淹没，能够高效率驾驭海量数据，获得有价值的决策信息。例如，在企业经营管理中，大数据能够帮助企业分析大量数据而进一步挖掘细分市场的机会，最终能够缩短企业产品研发时间，提升企业在商业模式、产品和服务上的创新力；学校和老师能够在对教学案例进行大数据分析的基础上改进他们的教学方法并合理安排教学内容；交管部门通过整合交通状况、天气以及驾驶员的地点信息等数据，可以更好地管理交通；大数据在政府和公共服务领域的应用，可以有效推动政府工作开展，提高政府部门的决策水平、服务效率和社会管理水平。

最后，通过大数据进行预测。《大数据时代》一书作者维克托·迈尔—舍恩伯格认为预测是大数据的核心，通过对大数据的分析来预测事情发生的可能性和发展的方向。例如，美国加州警方应用大数据进行预测分析，发现了犯罪趋势和犯罪模式，甚至可以对重点区域的犯罪概率进行预测；又如，前面提到的图灵奖获得者、著名数据库专家吉姆·格雷博士提出了第四种科学研究的方法——基于数据探索的方法，这种方法的本质就是基于大数据探索与发现自然和社会的规律。

大数据正日益对生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。大数据的应用已逐步深入我们生活的方方面面，涵盖医疗、交通、金融、教育、体育、零售等各行各业。下面我们列举几个大数据应用的典型案例。

（1）2014 年最热门的美剧非《纸牌屋》莫属。《纸牌屋》风靡北美乃至全球的一个重要原因，是大数据分析的结果。美国网飞（Netflix）公司是一家在线影片租赁提供商，该公司的网站有近 3 000 万订阅用户，这些用户在网站上收看视频的大量行为数据都被记录下来。据统计，用户每天在网飞上产生 3 000 多万个行为，包括暂停、回放、添加书签以及每天 300 万次搜索、400 万个评分。网飞对这些数据和收视调查等相关数据进行综合分析后发现，喜欢观看 BBC 老版《纸牌屋》的用户，大多喜欢大卫·芬奇导演或凯文·史派西主演的电视剧，于是网飞做出了拍摄《纸牌屋》的决策，投资 1 亿美元拍摄了新版《纸牌屋》，请大卫·芬奇执导、凯文·史派西做主演。结果，大数据技术让网飞公司赚得盆满钵溢。

（2）无论是在国内还是国外，体育行业都蕴含巨大的商机。例如，美国职业篮球联赛（NBA）的纽约尼克斯队在 2013 年就产生了 2.87 亿美元的收入。各支球队为了最大化自己的收入，必须在球场上不断赢球，因此教练组和相关人员必须一直做出正确的决策。而在这些决策中，体育的大数据分析扮演了一个日益重要的角色。

2015—2016 美国 NBA 赛季，骑士从 1∶3 落后，到 4∶3 夺冠，创造了 NBA 总决赛的历史。但球员的爆发，大劣势下的逆转，这一切的发生都不是偶然的。大数据文摘发现，在 2015—2016NBA 总决赛最后一场，骑士队的后卫 JR 史密斯在场上很好地充当了球队

第三得分点，13 投 5 中得到 12 分 4 篮板 2 助攻，其中三分球 8 中 2。这样的例子其实在 NBA 的赛场上比比皆是，球员并不是机器，他们的语言、行为其实都无时无刻不在透露大量可被分析和深度挖掘的信息。如何有效地将这些信息转化为知识，又如何利用这些知识来帮助人们做正确的决策？

运用大数据的体育数据分析包括运用统计工具来分析球员的历史表现。球队老板凭借分析结果来组建球队，教练组结合分析结果和他们的专业知识来调整上场阵容，提高球员的赛场表现。比如，利用非结构化社交媒体数据来提升现有体育分析模型效率，通过自然语言处理和文本挖掘技术来分析 NBA 球员的推文以了解他们的赛前情绪，从而提高对球员赛场表现的预测的准确性。

比如，2016 年 5 月 9 日西部半决赛第四场，雷霆主场战胜马刺，成功扳平大比分。而当地时间是母亲节，杜兰特全场出场 43 分钟，拿下 41 分，5 篮板，4 助攻，成为球队取胜的关键。众所周知，杜兰特与母亲感情非常好，其第一次荣获常规赛 MVP 发表演讲时，更是着重描述了童年时母亲的不易以及与母亲感情的深厚。而在比赛前，两队的明星球员中，只有杜兰特特意发表推文“So proud of my mama”，以此来表达对母亲的感谢，这也就不难解释杜兰特在本场比赛的爆发了。[①]

（3）2015 年 5 月，美国费城外一列美国铁路公司火车在一处急转弯路段发生脱轨事故，造成 5 人死亡和超过 200 人受伤。在费城到纽约的这一常用路段上，此次事故的发生显得非比寻常。次日早晨，半岛电视台美国频道发布了脱轨前火车的准确行驶速度：每小时 106 英里（约合每小时 170 千米），这超过了该路段限速（每小时 80 千米）的 2 倍之多。

之所以能如此迅速地做到这一点，是因为在此事发生的一年之前，他们就已经开始仔细调查美铁列车，设计了追踪其行驶的地图，每隔 5 分钟收集和存储一次数据。数据可以提供国内每列火车的实时定位和行驶速度。因此，通过找到事故发生之前的定位，他们在一张交互式的注释图中准确定位了该趟列车的行驶轨道。在后续追踪和分析从同一弯道通过的几百趟火车的行驶数据后，他们发现大部分火车的行驶速度都低于 50 英里/小时，而出事的火车却是一个特例。

该报道获得了 2016 年全球数据新闻奖（DJA）年度最佳突发新闻数据使用奖。

（4）淘宝目前占据中国网络购物 75% 的市场份额，每天产生的数据量达到了 7 T（7 000G）。这些数据当中大部分是由消费者、商家产生的交易数据，包括交易时间、商品价格、购买数量等，更重要的是，这些信息可以与客户和商家的年龄、性别、地址，甚至兴趣爱好等个人特征信息相匹配。阿里巴巴集团董事局主席马云表示，阿里巴巴公司本质上是一家数据公司，做淘宝不是为了卖货，而是为了获得所有零售的数据和制造业的数据；做物流不是为了送包裹，而是为了将这些数据融合在一起。淘宝数据魔方是淘宝网的大数据分析平台，通过这一平台，商家可以了解淘宝网上的行业宏观情况和自己品牌的市场状况，也可以分析竞争对手，探究消费买卖行为等，并据此进行生产、库存决策，而与此同时，更多的消费者也能以更优惠的价格买到更心仪的宝贝。另外，阿里信用贷款则是阿里巴巴通过所掌握的企业交易数据，借助大数据技术自动分析判定是否给予企业贷款，全程

① 该示例引自《大数据文摘》2016 年 6 月 21 日，“如何利用 NBA 球员推文预测其球场表现？”

只有少量人工干预。据透露，截至目前阿里巴巴已经放贷300多亿元，坏账率仅0.3%左右，大大低于同类银行。

(5)美国加州大学洛杉矶分校的研究者根据大数据理论设计了一款“电力地图”，将人口调查信息以及电力企业提供的用户实时用电信息与地理、气象等信息全部集合在一起，制作了一款加州地图。该地图以街区为单位，展示每个街区在当下时刻的用电量，甚至还可以将这个街区的用电量与该街区人的平均收入和建筑物类型等进行比照，从而得出更为准确的社会各群体的用电习惯信息。这个地图为城市和电网规划提供了直观有效的负荷数预测依据，知道哪些地区的用电负荷和停电频率过高，甚至可以预测哪些线路可能出现故障。

(6)UPS是总部位于美国亚特兰大的全球最大包裹快递公司，5个工作日在全球的送件量就能达到15.8亿件。为了监督管理员工并优化行车路线，UPS在货车上安装了GPS等传感器，由此获得了货车的各种运行数据，包括送货时间、行车路线、燃油消耗等，UPS采用其开发的Orion系统对这些海量数据进行道路优化分析。据报道，Orion可实时分析20万种可能路线，能在大约3秒内找出最佳路线。Orion的分析结果还表明卡车左转会导致货车长时间的等待。截至2013年年底，Orion已经在大约1万条线路上得到使用，这让UPS公司节省了150万加仑燃料，少排放了1.4万立方公吨的二氧化碳。

1.5 大数据时代的新机遇和新挑战

大数据时代，“资源”的含义正在发生极大的变化，它已不再仅仅只是指石油、煤、矿产等一些看得见、摸得着的实体，大数据也正在演变成为不可或缺的基础性战略资源。互联网和物联网每天都在产生大量的数据，这些庞大的数据资源为人类社会的发展提供了强大的推动力量。

1.5.1 依据大数据进行决策成为一种新的决策方式

从大数据中获取有价值的知识，让数据主导决策，是一种前所未有的决策方式。大数据分析和预测在人类决策管理方面正扮演着越来越重要的角色。例如，2009年美国爆发了甲型 H_1N_1 流感病毒，谷歌公司通过观察人们在网上搜索的大量记录，在流感爆发的几周前，就判断出流感是从哪里传播出来的，从而使公共卫生机构的官员获得了极有价值的数据信息，并做出有针对性的行动决策，而这比疾病控制中心的判断提前了两周。又如，美国的Farecast系统的一个功能就是飞机票价预测，通过分析从旅游网站获得的大量机票销售价格数据，预测出某一航班的机票价格在未来一段时间内的涨跌趋势，从而帮助乘客选择最佳的购票时机，从而降低购票成本。

1.5.2 大数据与各行业深度融合带来层出不穷的新应用

当今社会，政府、工业、交通、物流、商贸、金融、电信和能源等行业领域甚至新闻传媒领域都正在遭遇爆发式增长的数据量。据报道，美国很多世界500强大企业拥有大量的

数据，其平均拥有的数据量已经远远超过了美国国会图书馆所拥有的数据量。大数据的存在加速了各行各业与信息技术的深度融合。有专家指出，融合是大数据的价值所在，大数据与各行业深度融合会进一步释放大数据的能量，从而改变当今社会每一个行业的管理模式和生产经营模式。例如，在农业领域，硅谷有一家气候公司，从美国气象局等数据库中获得几十年的天气数据，将各地降雨、气温、土壤状况与历年农作物产量的相关度做成精密图表，预测农场来年产量，向农户出售个性化保险。在商业领域，沃尔玛公司通过分析销售数据，了解顾客购物习惯，得出适合搭配在一起出售的商品，还可从中细分顾客群体，提供个性化服务。在金融领域，华尔街德温特资本市场公司分析 3.4 亿微博账户留言，判断民众情绪，依据人们高兴时买股票、焦虑时抛售股票的规律，决定公司股票的买入或卖出。在社会安全管理领域，通过对手机数据的挖掘，可以分析实时动态的流动人口来源、出行，实时交通客流信息及拥堵情况。利用短信、微博、微信和搜索引擎，可以收集热点事件，挖掘舆情，还可以追踪造谣信息的源头。美国麻省理工学院通过对十万多人手机的通话、短信和空间位置等信息进行处理，提取人们行为的时空规律性，进行犯罪预测。

1.5.3　大数据推动新技术的不断涌现

数据科学与其他学科的融合以及大数据的应用需求，导致了新学科和新技术的不断涌现。例如，在科学研究领域，基于密集数据分析成为继实验科学、理论科学和计算科学之后的第四种科学探索方式，基于大数据分析的材料基因组学和合成生物学等正在兴起。大数据带动了数据可视化分析技术的研究和发展，利用计算机自动化分析能力的同时，充分挖掘人对于可视化信息的认知能力优势，将人、机的各自强项进行有机融合，借助人机交互式分析方法和交互技术，辅助人们更为直观和高效地洞悉大数据背后的信息、知识与智慧。近年来，大数据与神经计算、深度学习、语义计算以及其他人工智能相关技术相结合，促进人工智能技术不断提高，使计算机系统拥有了更好的对数据的理解、推理、发现和决策能力。

目前，虽然社会上出现了一些应用大数据技术的成功案例，但是大数据的应用仍存在一些困难与挑战，主要体现在以下四个方面。

第一，在数据收集方面。大数据的数据量不仅巨大，而且数据结构种类繁多，不仅仅有简单的、结构化的数据，更多的则是复杂的、非结构化的数据，而且数据之间的关系较为复杂。如何从不同的数据源及时收集到所需要的数据面临巨大的困难，并且大量不同数据源的数据之间可能存在冲突、不一致或相互矛盾的现象。为了保证所收集的数据的质量，就必须识别和检测大数据中的错误、缺失、无效数据，这给大数据环境中数据质量的监测和管理带来巨大的挑战。

第二，在数据存储方面。由于大数据的数据结构的多样性，单一的数据结构（如传统关系型数据库中的二维表结构）已经远远不能满足大数据存储的需要。据调查，目前国内外大部分企业的业务数据仍以结构化数据为主，相应地主要采用传统关系型数据库进行数据的存储。对于非结构化数据，则是先将其转化为结构化数据后再进行存储、处理及分析。这种数据存储处理方式不仅无法应对大数据数量庞大、数据结构复杂、变化速度快等特点，而且一旦转化方式不当，将会直接影响数据的完整性、有效性与准确性。因此，需

要开发专门的数据库技术和专用的数据存储设备进行大数据的存储，保证数据存储的有效性。

第三，在数据分析处理方面。有些行业的数据涉及上百个参数，其复杂性不仅体现在数据本身，更体现在数据之间在多源异构、多实体和多空间之间的动态关联，难以用传统的方法描述与度量，处理的复杂度很高。例如，需要将高维图像等多媒体数据降维后度量与处理，或者利用上下文关联进行语义分析，从大量动态而且可能是模棱两可的数据中综合信息，并导出可理解的内容等。目前，尽管计算机智能分析技术有了很大进步，但还只能针对小规模、有结构或类结构的数据进行分析，还不能胜任对大数据的深层次的数据挖掘。另外，速度是规模的另一面，需要处理的数据集越大，分析所花费的时间将越长。在大数据背景下，许多时候面对汹涌的数据流要求立即得到分析结果，这种及时性的要求也是大数据分析处理的另一个挑战。

第四，在安全风险方面。首先，大数据容易成为黑客攻击的首要目标。大数据是宝贵的资源，不仅意味着海量的数据，也意味着更复杂、更敏感的数据，这些数据会吸引更多的黑客，成为更具吸引力的目标。并且大数据中的数据大量聚集，使得黑客一次成功的攻击就能导致严重的安全事故，例如，用户大量的个人信息被泄露。其次，大数据加大了隐私泄露风险。大数据的来源涵盖非常广阔的范围，如可能来自可穿戴设备的传感器、社交网络、智能手机、电子邮件等，这些数据可能包含了个人的隐私和各种行为的细节记录，大量个人数据的聚集不可避免地加大了隐私泄露的风险。

1.6 本书的特定视野

大数据是信息社会的宝贵资源。但大数据的价值是掩埋在沙子之中的，大数据的价值也不是单方面的，而是多元的。要把大数据的价值发掘出来，就要凭借数据分析。数据分析有多种类型，适用于多种目的。抱着不同的目的，运用不同的方法，我们可以在多个方面对大数据进行不同的分析。在统计学领域，有些人将数据分析划分为描述性统计分析、探索性数据分析和验证性数据分析。所谓描述性统计分析，就是对一组数据的各种特征进行分析，以便描述测量样本的各种特征及其所代表的总体的特征。描述性统计分析的项目很多，常用的如平均数、标准差、中位数、频数分布、正态或偏态程度等。这些分析是复杂统计分析的基础。所谓探索性数据分析（exploratory data analysis，EDA）是指对已有的数据（特别是调查或观察得来的原始数据）在尽量少的先验假定下进行探索，通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。特别是当我们对这些数据中的信息没有足够的经验，不知道该用何种传统统计方法进行分析时，探索性数据分析就会非常有效。探索性数据分析在 20 世纪 60 年代被提出，其方法由美国著名统计学家约翰·图基（John Tukey）命名。验证性数据分析则侧重已有假设的证实或证伪，在应用分析中，大量的分析都涉及验证性数据分析。本书所列举的案例，大多数涉及的就是验证性数据分析。大数据是浩瀚无际的大海，我们每个人现在有能力采撷的可能只是小小的一杯水，或者一朵小小的浪花。本书无论从案例的分析或是其他方面的论述，侧重的都是大数据在经济社会实际应用中的求证或求伪，这是本书的特定

视野。本书选择的都是在政府服务、经济监督、社会治理和经济运行等方面提出的实际问题，探索用大数据分析的方法回答和处理，也就是坚持问题导向，把大数据分析与经济社会的实际应用需求紧密结合起来，在融合中探索大数据解决方案。

大数据和传统数据库有着密切的联系。传统数据库是大数据的一个重要的组成部分，大数据是传统数据库处理能力的拓展和延伸。从这种认识出发，本书在探索数据分析技术和方法，分析应用案例的时候既注意敏锐观察大数据发展的新特点，又注意运用传统的数据分析技术，尤其注意不同数据的融合，综合运用适用的技术和方法。这也是本书讨论应用分析技术和方法的一个着重点。

参考文献

[1] 李学龙，龚海刚. 大数据系统综述[J]. 中国科学：信息科学，2015，(1)：1-44.

[2] [英]维克托·迈尔—舍恩伯格，肯尼思·库克耶. 大数据时代[M]. 盛杨燕，周涛，译. 杭州：浙江人民出版社，2013.

[3] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，(1)：146-169.

[4] 刘智慧，张泉灵. 大数据技术研究综述[J]. 浙江大学学报：工学版，2014，(6)：957-972.

[5] 覃雄派，王会举，杜小勇，王珊. 大数据分析——RDBMS 与 MapReduce 的竞争与共生[J]. 软件学报，2012，23(1)：32-45.

第 2 章　大数据应用分析

2.1　大数据的处理流程

当生产和生活中产生了海量数据之后，人们为了充分发现和利用这些海量数据蕴藏的价值，需要对这些海量数据进行一系列的处理。大数据的处理流程可以定义为在合适工具的辅助下，对大量异构的数据源进行抽取和集成，然后按照一定的标准统一存储，再利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识并利用恰当的方式将结果展现给终端用户。大数据的数据来源广泛，由此导致应用需求和数据类型千差万别，但总的来说，大数据的基本处理流程大都是一致的。大数据处理流程具体可划分为数据采集、数据的处理与集成、数据分析和数据的解释四个阶段，如图 2-1 所示。

整个大数据的处理流程大致如下：首先，从大量异构的数据源获取数据；其次，根据数据类型的不同（包括结构化数据、半结构化数据和非结构化数据），采用特殊方法（包括数据聚合、数据修正、数据清洗、数据去噪）对数据进行处理和集成，将其转变为统一标准的数据格式；最后，用合适的数据分析方法对这些数据进行分析，并利用可视化等技术将分析的结果展现给用户。

1. 数据采集

数据采集是大数据处理流程中最基础的一步，由于大数据的数据量大、数据种类复杂，因此，通过各种方法获取数据便显得格外重要。目前常用的数据采集手段包括通过传感器（如麦克风、摄像头等）获得、通过读取射频识别卡（RFID）信息、通过搜索引擎（如百度和谷歌等）采集以及通过网页爬虫从互联网上采集等。随着物联网、移动设备和社交网络的普及和发展，所需采集的数据量会变得越来越大，数据类型也会千差万别。

2. 数据的处理与集成

数据的处理与集成主要是对已经采集到的数据进行适当的处理，清洗去噪以及进一步集成存储。从第 1 章中，我们知道数据种类繁多是大数据的一个重要特点，这就决定了从各种渠道获取的数据种类和结构都非常复杂，给之后的数据分析处理带来了极大的困难。通过数据的处理与集成这一步骤，将这些结构复杂的数据转换为单一的或是便于处理的数据结构，为以后的数据分析打下良好的基础。因为这些数据里并不是所有的信息都是必需的，而是会掺杂很多噪声和干扰项，因此，还需对这些数据进行“去噪”和清洗，以保证数据的质量以及可靠性。

3. 数据分析

数据分析是整个大数据处理流程里核心的部分，因为大数据的价值产生于分析过程。在数据分析的过程中，会发现数据的价值所在。经过上一步骤数据的处理与集成后，所得

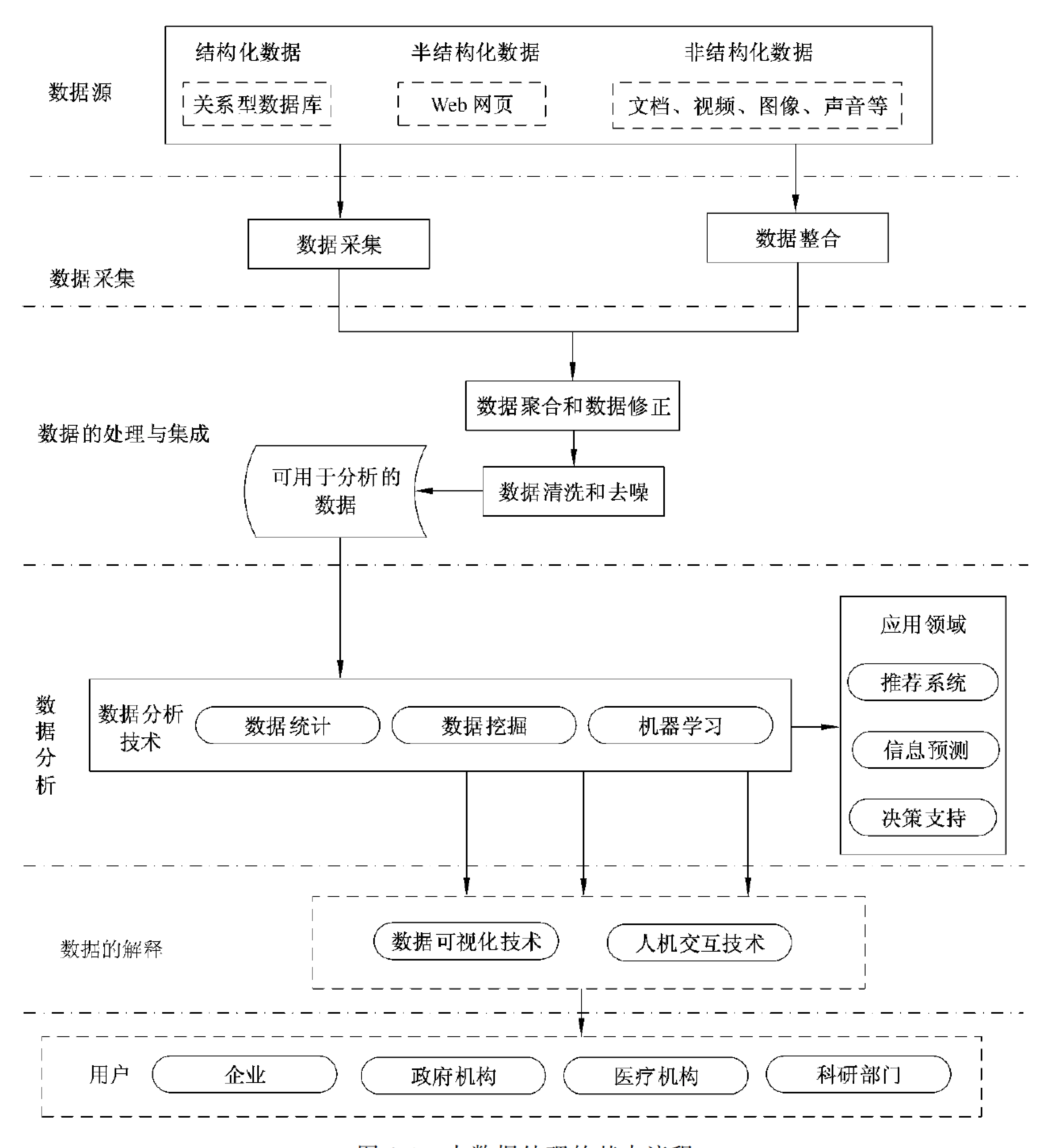


图 2-1　大数据处理的基本流程

到的数据便成为数据分析的原始数据，用户会根据不同的分析目的和应用需求对这些数据进行进一步的处理和分析。数据分析方法主要包括数据挖掘、机器学习、智能算法、统计分析等。本章首先重点介绍大数据分析的基本概念，最后将通过具体案例进一步介绍大数据分析的方法。

4. 数据的解释

对于广大的数据用户来讲，最关心的并非数据分析处理的过程，而是对大数据分析结果的解释与展示。因此，在一个完善的数据分析流程中，对数据分析结果的解释至关重要。若数据分析的结果不能得到恰当的显示，则会对数据用户产生困扰，甚至会误导用户。传统的数据显示方式是以文本形式或者简单图形显示结果。这种方法在数据量小时是一种很好的选择，但是大数据时代的数据是海量的，同时数据之间的关联关系极其复

杂,采用传统的解释方法不能直观、明白地向用户展示数据蕴含的规律。因此,现在人们引入了"数据可视化技术"来展示和解释大数据的分析结果。通过可视化技术,可以形象地向用户展示数据分析结果,更方便用户对结果的理解和接受。另外,人们还采用人机交互技术,利用交互式的数据分析过程来引导用户逐步地进行分析,使用户在得到结果的同时更好地理解分析结果。

2.2　大数据分析的概念

数据分析是指用适当的统计方法对收集来的大量数据进行分析,为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。在实际应用中,数据分析可以帮助人们作出判断,以便采取合适的行动或措施。数据分析的数学基础在 20 世纪早期就已确立,但直到计算机的出现才使实际操作成为可能,并使数据分析得以推广。数据分析是数学与计算机科学相结合的产物。

大数据分析是指对规模巨大的数据进行分析。大数据之所以备受关注,本质原因在于其具有巨大的潜在价值。大数据分析技术作为获取数据价值的关键手段,在大数据应用中占有极其重要的位置,可以说是决定大数据价值能否发掘出来的关键因素。数据分析是整个大数据处理流程的核心。在数据分析过程中,人们采用适当的方法(包括统计分析和数据挖掘等方法),对采集到的海量数据进行详细研究和概括总结,从而发现和利用其中蕴含的信息和规律。大数据分析的主要目标包括:推测或解释数据、检查数据是否合法、给决策提供合理建议、诊断或推断错误原因以及预测未来将要发生的事情。

根据大数据的数据类型,可以把大数据分析划分成如下三类。

(1) 结构化数据分析:对传统关系数据库数据的分析。

(2) 半结构化数据分析:对 HTML 网页或 XML 文档等半结构化数据的分析。

(3) 非结构化数据分析:对图像、声音和视频等非结构化数据的分析。

值得一提的是,大数据时代,相关分析因其具有可以快捷、高效地发现事物间内在关联的优势而受到广泛的关注。所谓大数据相关关系,是指 2 个或 2 个以上因素之间在某种意义下所存在的联系和规律。相关分析的目的在于探寻大数据集里所隐藏的内在关联关系。近年来大数据相关分析的应用成果不断涌现,人们日益发现,和以往相比,大数据时代相关关系的探索具有更加重要的价值。例如,在电子商务推荐系统中,通过挖掘用户性别、家庭情况、居住位置、以往的购物情况、商品特性之间的相关关系,能够进行有针对性的商品推荐。又如,商业企业作为大数据应用的重要领域,通过分析管理措施和经营策略与利润增长具有何种相关性,可以帮助企业管理者调整经营策略,实现企业利润的增长。综合来看,大数据相关分析已经成为大数据分析与挖掘的核心科学问题和关键应用技术。

大数据分析的出现不是对传统数据分析的否定,而是对传统数据分析的继承和发展,传统数据分析方法中的数据挖掘和统计分析仍然在大数据分析中发挥重要的作用。同时,大数据分析也呈现出和传统数据分析不同的特征,表现在如下四个方面。

第一,所分析的数据量不一样。传统的数据分析是对少量的数据样本进行分析,而正

如著名的大数据专家迈尔—舍恩伯格在其名著《大数据时代》一书中指出的：大数据要分析的是与某事物相关的所有数据，而不是依靠分析少量的数据样本。

第二，分析的侧重点不一样。迈尔—舍恩伯格在《大数据时代》一书中指出：大数据分析的重点不是发现事物之间的因果关系，而是发现事物之间的相关关系，因此相关分析是大数据分析的重要内容。

第三，所分析数据的来源不一样。传统数据分析的对象大多局限在同一个来源的数据中，如 Oracle 数据库或者 SQL Server 数据库中的数据，但是大数据分析更强调数据融合，因为每一种数据来源都有一定的局限性和片面性，只有对各种来源的原始数据进行融合才能反映事物的全貌。事物的本质和规律往往隐藏在各种原始数据的相互关联之中。

第四，数据的解释方式不一样。可视化分析在传统数据分析中只是一种辅助分析手段，但是大数据分析中更强调可视化分析的应用。俗话说“一幅图胜过千言万语”，大数据的数据内容纷繁复杂，可视化分析能够直观地呈现大数据的特点，有利于用户发现和掌握其中的规律。

2.3　大数据分析的关键技术

大数据分析是挖掘大数据价值的手段，大数据分析技术对于准确、高效获得大数据中隐藏的模式和规律至关重要。目前大数据分析的关键技术包括云计算、数据分析和可视化等多种技术，这些方法随着大数据的发展，其内涵和外延也在不断发展和变化。

2.3.1　云计算

云计算是大数据分析处理的基础，也是大数据分析的支撑技术。如果将各种大数据的应用比作一辆辆“汽车”，支撑起这些“汽车”运行的“高速公路”就是云计算。正是云计算技术在数据存储、管理与分析等方面的支撑，才使得大数据有了广阔的用武之地。因此，在大数据时代，大数据是需求，云计算是手段，没有云计算就无法处理大数据。

对于云计算，美国国家标准与技术研究院的定义是：“云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，让使用者可以访问可配置的计算资源（资源包括网络、服务器、存储、应用软件、服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。”国内专家对云计算给出了更加简洁的定义：“云计算是一种商业计算模型。它将计算任务分布在异地大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和信息服务。”在这个定义中，提供资源的网络被称为“云”，这些资源包括计算服务器、存储服务器和网络带宽资源等。“云”通过网络向使用者按需提供可动态扩展的廉价计算服务和存储服务。“云”中的资源在使用者看来是可以无限扩展的，并且可以随时获取、按需使用、随时扩展、按使用付费。通过云计算，使用者不需要购买昂贵的硬件设备和操作软件，也不需要专门的 IT 维护人员，只需要通过网络就可以随时随地使用云计算强大的计算能力。因此，有人将云计算比喻为从单台发电机供电模式转向了电厂集中供电的模式，这意味着计算能力也可以作为

一种商品进行流通，就像水、电和煤气一样，取用方便，费用低廉。

云计算按照服务类型大致可以分成如下三类。

(1) 将基础设施作为服务(IaaS)：IaaS将硬件设备等基础资源封装成服务提供给用户使用。例如，亚马逊公司的弹性计算云EC2和简单存储服务S3就是IaaS的典型代表。在IaaS中用户相当于获得了裸机和磁盘，可以任意安装所需要的软件，如可以安装Windows和MS Office。

(2) 将平台作为服务(PaaS)：PaaS提供用户应用程序的运行环境。典型的PaaS包括谷歌的App Engine和微软的MS Windows Azure。PaaS相当于给用户提供一台安装了操作系统的计算机，用户可以在这个平台上继续安装所需要的其他软件。

(3) 将软件作为服务(SaaS)：SaaS将某些特定的应用软件封装成服务提供给用户，用户不需要在本地安装这些软件，只需通过SaaS就可以在线使用。例如，谷歌的在线文档处理软件Google Docs就是典型的SaaS。用户不需要在本地PC机上安装这个软件，只要有网络，就可以在线使用它来完成文档编辑、排版、保存和打印的工作。

为了处理海量的Web数据，谷歌于2006年首先提出了云计算的概念。谷歌基于云计算平台开发了支持大数据应用的一系列技术，包括分布式文件系统GFS、分布式数据处理MapReduce以及分布式数据库Bigtable。这些技术获得了广泛的应用，其中GFS为整个大数据提供了底层的数据贮存支撑架构，GFS能够处理的文件很大，容量通常都在100MB以上，而且大文件在GFS中可以被有效地管理；MapReduce是一种处理海量数据的并行运算模式，特别适用于非结构化和结构化的海量数据的搜索、挖掘和分析；Bigtable是非关系型数据库，能够有效存储和管理大数据中的半结构化数据和非结构化数据，这对大数据集中占较大比例的非结构化数据非常适用。

这些技术对大数据的分析处理产生了深远影响，催生出以Hadoop为代表的一系列开源大数据处理工具。

2.3.2 数据分析方法

尽管目标和应用领域不同，一些常用的分析方法(如统计分析和数据挖掘)对大数据同样适用。

(1) 统计分析：在统计理论中，通过概率理论对数据的随机性和不确定性建立模型。统计分析技术可以分为描述性统计技术和推断性统计技术。描述性统计技术对数据集进行总结或描述，而推断性统计技术则能够对过程进行推断。统计分析方法包括回归、因子分析、聚类和判别分析等。

(2) 数据挖掘：数据挖掘是发现大数据集中数据模式的计算方法。许多数据挖掘算法已经在人工智能、机器学习、模式识别、统计和数据库领域得到了广泛应用。著名的数据挖掘算法包括决策树、k-means算法、支持向量机、Apriori算法、最大期望算法、PageRank算法、AdaBoost算法、k最邻近算法、朴素贝叶斯和分类与回归树，覆盖了分类、聚类、回归和统计学习等方面。另外，深度学习和遗传算法等先进的智能技术也被用于数据挖掘中。

2.3.3　数据可视化

数据可视化是解释大数据的有效手段之一。数据可视化技术(Data Visualization)是指运用计算机图形学和图像处理技术,将数据转换为图形或图像在屏幕上显示出来,并进行交互处理的理论、方法和技术。图形化的方式比文字更容易被用户理解和接受,数据可视化将抽象的数据表现成为可见的图形或图像,帮助人们直观形象地发现数据中隐藏的内在规律。

一般来说,图表和地图可以帮助人们快速理解信息。但是,当数据量增大到大数据级别时,传统的电子表格等技术已不能清晰展现海量数据的特点,因此需要研究适用于大数据的可视化手段。目前,大数据的可视化已成为学术界和工业界的一个活跃的研究领域,出现了一些成功的案例。例如,大众点评网通过地图的方式向用户呈现每一天全国各地餐厅最火的菜品以及人均消费,为用户的消费提供参考。又如,支付宝每隔一段时间(通常为一个月)产生用户的可视化对账单,其中反映了用户所在地区的消费趋势以及用户本人的消费情况和偏好,帮助用户管理自己的消费支出,如图 2-2 所示。

另外,也出现了一些支持数据可视化的工具和软件,如 R 语言和商业数据分析软件 Tableau 都提供了强大的数据可视化分析功能。

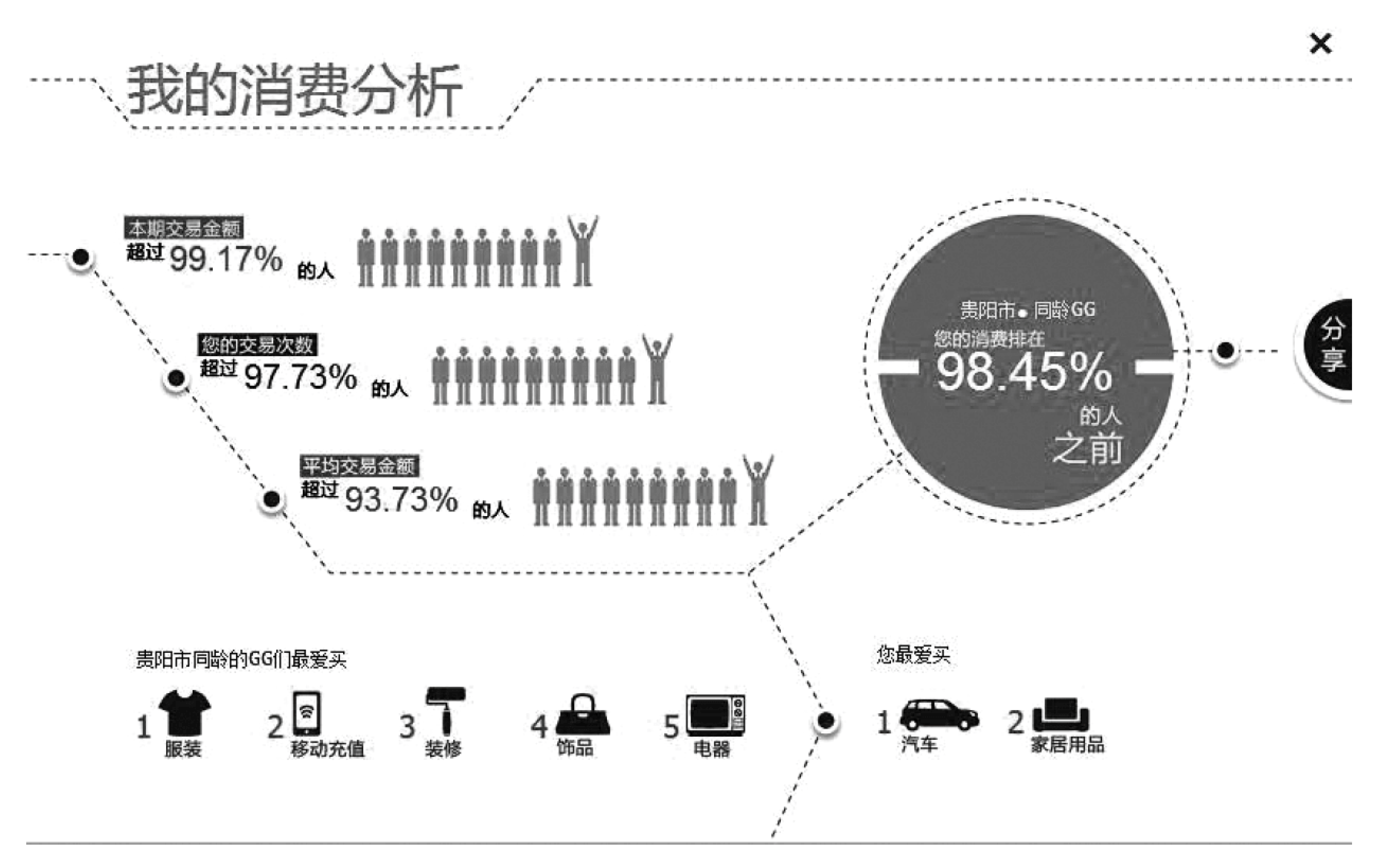


图 2-2　支付宝对账单的可视化分析

2.4　大数据分析工具介绍

随着大数据应用的不断发展和普及,诞生了多种类型的大数据分析工具,这些工具各有千秋,有的偏重数据分析的效率和扩展的灵活性,有的偏重可视化分析,有的只适用于

特定的应用领域。下面简单介绍五种目前最流行的大数据分析工具。

2.4.1 Hadoop

在大数据时代,许多传统的数据分析技术和数据库技术已经不足以满足需求,许多大数据应用对于数据分析和管理都提出了新的要求。正如前面提到的,云计算是大数据分析处理的关键技术,谷歌的云计算技术对大数据的分析处理产生了深远影响。Hadoop是开源的云计算平台,它模仿和实现了谷歌云计算的主要技术。现在 Hadoop 已经发展为一个包括分布式文件系统 HDFS、分布式数据库 HBase 以及数据分析处理 MapReduce 等功能模块在内的完整生态系统,目前已经成为最流行的大数据处理平台。用户可以从 hadoop. apache. org 免费下载和安装 Hadoop 的相关软件。

英特尔公司根据大数据处理的要求,提出了一种 Hadoop 的组件结构,集中展现了大数据的采集、存储和分析处理的主要功能模块,如图 2-3 所示。

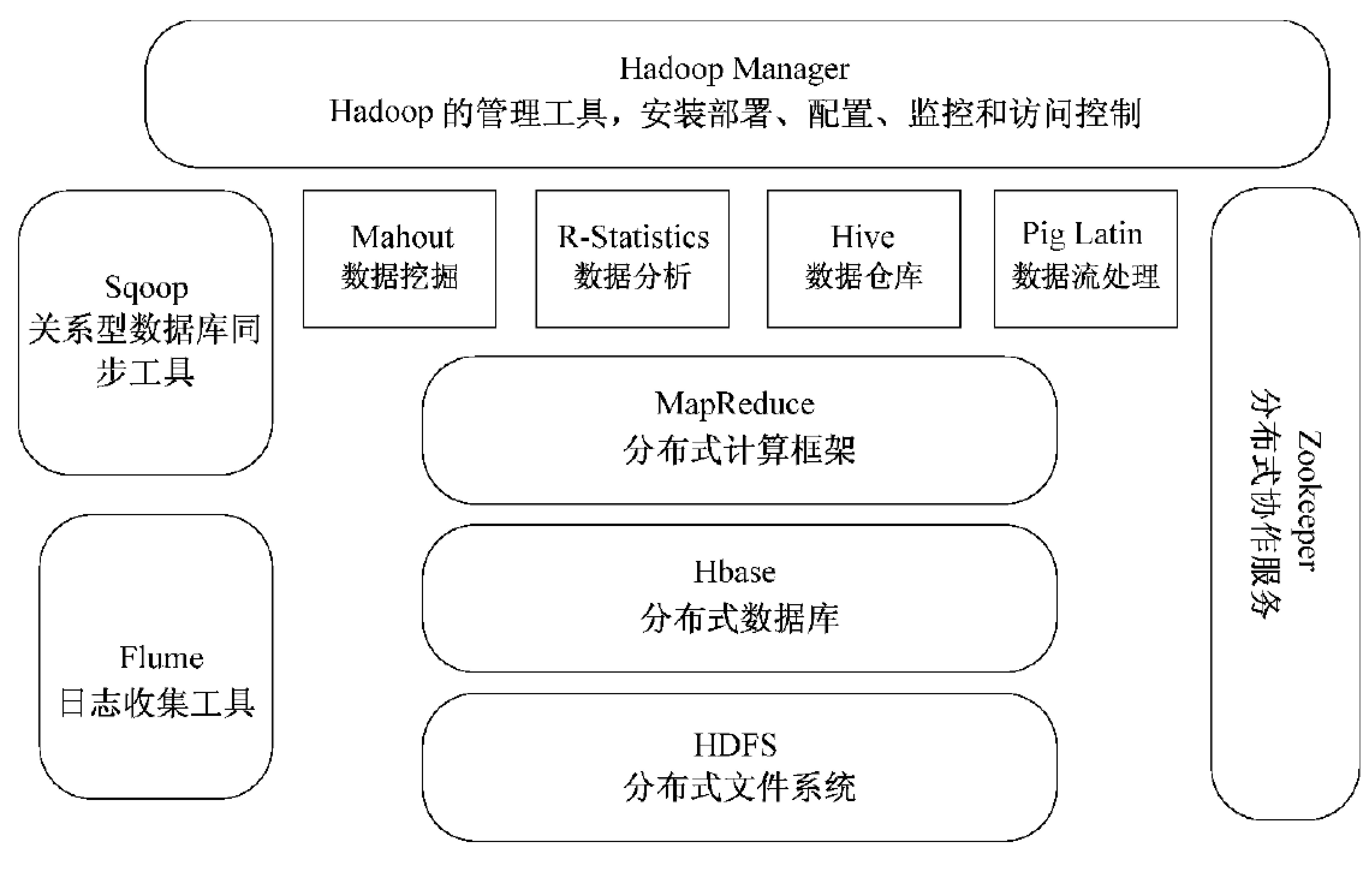


图 2-3 英特尔的 Hadoop 组件结构

在这个组件结构中,MapReduce 是分布式数据处理模式,它可以将复杂的处理任务分配给一群服务器,适合海量数据的处理。HDFS 是一种类似于谷歌 GFS 的分布式文件系统,可以为大规模的服务器集群提供高速度的文件读写访问。HBase 是一种与谷歌 BigTable 类似的分布式并行数据库系统,可以提供海量数据的存储和读写,并且兼容各种结构化或非结构化的数据。Mahout 是 Apache 软件基金会的一个开源项目,是对海量数据进行挖掘分析的软件,提供了丰富的数据挖掘和机器学习功能。Hive 是一种基于 Hadoop 的大数据分布式数据仓库,它将数据存储在相应的分布式数据库或分布式文件系统中,使用 SQL 语言对海量数据信息进行统计、查询和分析等操作。Pig Latin 是对大规模数据进行分析处理的语言,它结合了 SQL 和 MapReduce 两者的优点,可以像 SQL 语言那样灵活可变。Zookeeper 是分布式系统的协调系统,可以提供包括配置维护、名字服务、分布式同步、组服务等在内的相关功能。Sqoop 是一个用来将 Hadoop 和关系型数

据库中的数据双向转移的工具，可以将一个关系型数据库(如 Oracle、MySQL 等)中的数据导入 Hadoop 的分布式文件系统 HDFS 中，也可以将 HDFS 的数据导入关系型数据库中，还可以在传输过程中实现数据转换等功能。Flume 是一种分布式日志采集系统，它的作用是从不同的数据源系统中采集和传输大量的日志数据到一个集中式数据存储器中。

2.4.2　R

R 是当今最受欢迎的数据分析和可视化平台之一。R 是开源软件，任何人都可以从 www.r-project.org 免费下载和安装 R 软件。R 是由一种名为 S 的统计软件演变而来的。S 于 20 世纪 70 年代诞生在美国的贝尔实验室，由里克 · 贝克(Rick Becker)、约翰 · 钱伯斯(John Chambers)和艾伦 · 韦尔克斯(Allan Wilks)三人共同开发。1995 年新西兰奥克兰大学的罗斯 · 伊哈卡(Ross Ihaka)和罗伯特 · 杰特曼(Robertt Gentleman)重新实现了 S 的部分功能，并把所有源代码公开，这就是 R 软件。R 软件中的命令统称为 R 语言。

R 软件能够成为受人们欢迎的数据分析软件，是与其优秀的特性分不开的，主要包括下列特性。

(1) R 拥有强大的统计分析功能：R 内嵌了许多实用的统计分析函数，使用者可以轻松完成各种统计工作。除了 R 内嵌的统计函数外，R 还以“包”的形式提供了扩展的数据分析功能，以满足各种分析的需要。目前 R 拥有超过 2 100 种包，涵盖了基本统计、经济学、社会学、生态学、地理学、医学、生物信息等多个领域，使用者可以根据需要下载安装这些包来扩展 R 的统计分析功能。因此，几乎所有的统计分析工作都可以通过 R 来完成。

(2) R 拥有强大的可视化功能：R 提供了丰富的 2D 和 3D 绘图函数来完成数据可视化，并能将这些可视化结果保存为多种形式的文件，如 jpg、bmp、pdf、png 等。

(3) R 完全免费：用户可以免费下载使用。

(4) R 支持多种操作系统平台：R 可以运行在多种操作系统平台上，如 Windows、Linux、Unix 和 MacOS，当今主流的计算机平台都可以运行 R。

(5) R 的帮助功能完善：R 包含了一个非常实用的帮助系统，软件的帮助文件可以随时通过主菜单浏览和打印。通过 help 命令可以随时了解 R 所提供的各种函数的使用方法和例子。

2.4.3　Python

Python 是目前十分流行的编程语言，根据 2015 年 TIOBE 编程语言排行榜，Python 语言已成为除 Java 和 C/C++ 外，最受人欢迎的编程语言。Python 也是开源免费软件，用户可以从 www.python.org 下载和安装 Python 的开发平台。Python 语法简单清晰，容易学习掌握。Python 具有丰富和强大的库。它常被昵称为胶水语言，能够很轻松地把其他语言开发的各种功能模块集成到所开发的程序中。随着 Python 提供的统计分析和可视化函数库的不断增加和完善，Python 语言正成为一种数据分析的强大语言。Python 语言数据分析的函数库主要包括 NumPy、SciPy、IPython 和 Pandas，可视化函数库包括 Matplotlib，用户可以调用这些函数库完成各种数据分析和可视化任务。

2.4.4 RapidMiner

RapidMiner 是数据挖掘、机器学习和商业预测分析的开源软件，用户可以从 www.rapidminer.com 免费下载和使用。RapidMiner 除了内嵌数据挖掘和机器学习功能外，还可以与 R 软件进行协同工作，通过 R 扩展它的数据分析功能。RapidMiner 和 R 及 Python 最大的区别在于它不需要任何编程，只需通过鼠标拖放，就能完成数据挖掘和分析的功能。

在 RapidMiner 中整个数据挖掘过程就像是车间的生产流水线。在 RapidMiner 中输入原始数据，经过一系列流程后输出数据分析结果或预测结果。其中，流程(process)是指按照一定先后次序依次执行的一系列分析处理函数(称为算子，operataor)。不同算子有不同的输入输出特性。RapidMiner 大概包括以下几类算子：流程控制类，实现循环和条件功能；数据输入和输出类，实现数据交换；数据转换类，数据抽取、清洗整理功能；建模类，分类回归建模、关联分析、聚类分析、集成学习等；评估类，多重交叉检验、自助法检验等。

2.4.5 Tableau

Tableau 是当前最受欢迎的数据分析和可视化软件，在市场分析公司 Gartner 2015 年 2 月公布的商业智能分析平台的报告中，连续第三次蝉联领先者的殊荣。Gartner 在报告中指出：“Tableau 在简单易用方面是现有的商业智能分析软件中做得最好的。”德国的数据科学家 Lucie Salwiczek 也认为：“不管是制作报表，还是深入挖掘数据并进行分析，只需要 Tableau 这样一个工具就足够了。”

Tableau 之所以受到市场的欢迎，原因在于以下几个方面的主要特性。

- 简单易用：Tableau 提供了非常友好的可视化界面。用户不需要编写程序代码，只需通过点击鼠标和简单拖放，就可以迅速创建出精美直观和具有交互功能的报表、仪表盘、故事，帮助用户迅速发现和展示数据中的特征和规律。其操作非常简单，使用者不需要太多的 IT 背景和统计知识。
- 强大的可视化技术：可视化技术是 Tableau 的核心。Tableau 提供了一个非常新颖和简洁易用的操作界面，使用户在处理规模巨大的多维数据时，可以从不同角度和设置看到数据所呈现的规律，其自动生成的图表，既能准确反映数据的特征，也丝毫不逊色于专业美术编辑的水平，如图 2-4 所示。图中反映了 2015 年 4 月 30 日尼泊尔地震的多维分析，包括不同区域的地震级别和震源深度、不同区域的伤亡情况、外国人的伤亡情况、国际人道主义救援的情况等。Tableau 提供数据可视化技术，使数据挖掘变得简单易用，直观清晰。正是因为这个特点，Tableau 获得了数据分析专家的广泛认可，其用户数量逐年递增。
- 可连接多种数据源，轻松实现数据融合：在日常工作中，用户想要分析的数据可能分散在多个数据源中，有的存在于文件中，有的可能保存在数据库里面。Tableau 允许从多个数据源访问数据，包括文本文件、Excel 文件、Oracle 数据库、SQL 数据库和 Hadoop 数据文件等。Tableau 允许用户查看多个数据源，不仅能

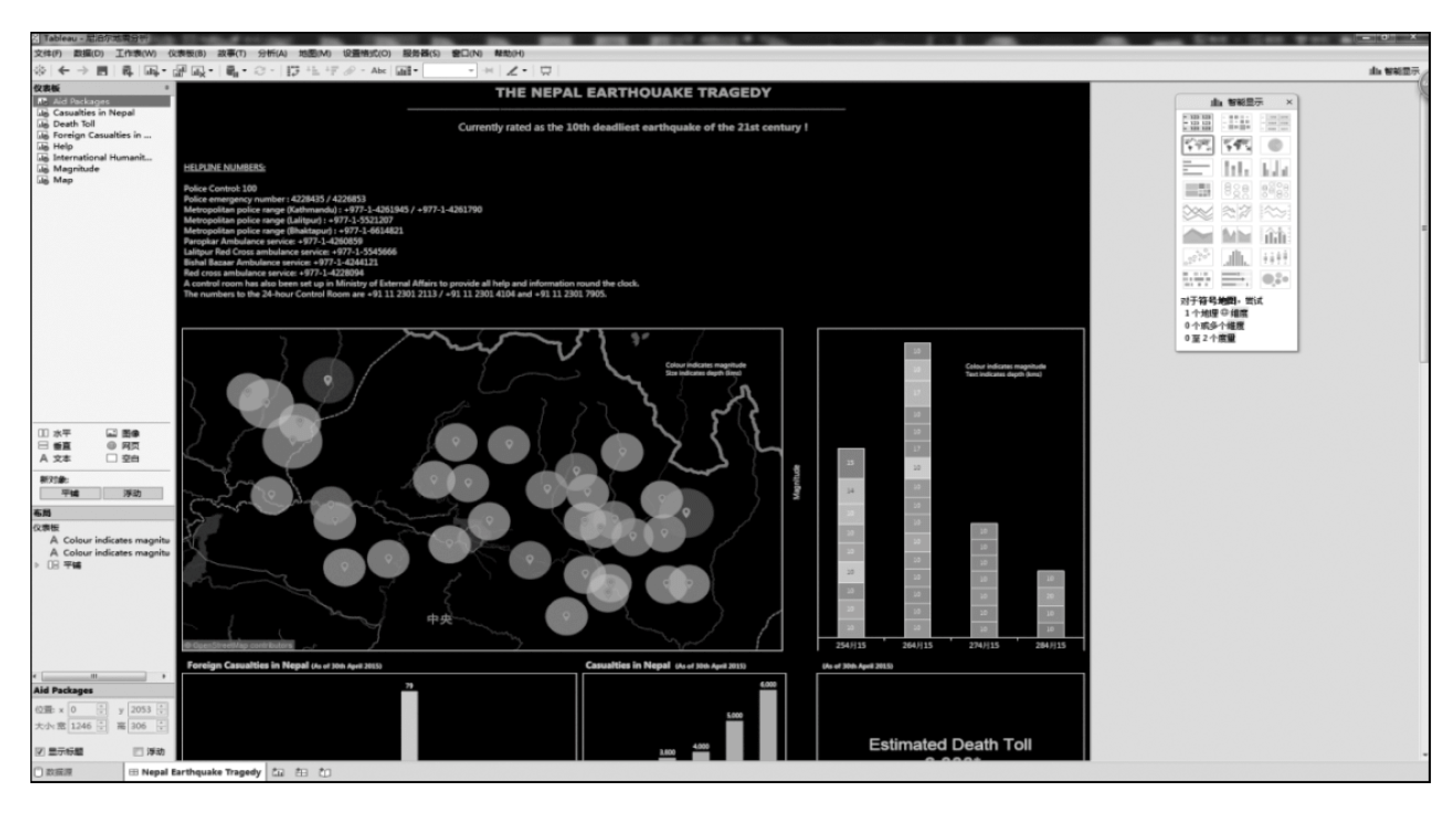


图 2-4　尼泊尔地震情况的多维度分析

够在不同数据源之间来回切换分析，也能够把多个不同数据源结合起来使用。

- 具有良好的可扩展性：Tableau 提供了多种应用编程接口来扩展其数据分析的能力，具体包括：通过数据提取接口可以连接使用多种格式的数据源；通过页面集成接口，把 Tableau 制作的报表和可视化内容嵌入已有的信息化系统或者商务智能平台中，实现与网页的集成和交互；通过与 R 的接口，充分利用 R 语言强大的统计分析和数据挖掘功能，提升 Tableau 在数据处理和高级分析方面的能力。

Tableau 的产品系列非常丰富，涵盖了从移动终端到企业级服务器的数据分析需求，具体包括 Tableau Desktop、Tableau Server、Tableau Online、Tableau Mobile、Tableau Public 和 Tableau Reader。表 2-1 对 Tableau 的各产品进行了简单介绍。

表 2-1　Tableau 系列产品的功能简介

产品名称	简 要 介 绍
Tableau Desktop	Tableau 的桌面分析软件
Tableau Server	Tableau 的企业级分析平台，可以发布和共享不同 Tableau Desktop 的分析结果，也可以发布和管理数据源
Tableau Online	基于云计算的数据分析平台，提供 Tableau Server 的所有功能，免去硬件和软件部署与维护，用户按照每人每年的方式付费使用
Tableau Mobile	是针对 iOS 和安卓平台的移动端分析程序。用户可以通过 iPad 或手机等移动设备来查看 Tableau Server 或 Tableau Online 上的分析结果，并可以进行简单的分析和编辑
Tableau Public	是一款免费的服务器软件，用户可以用它在互联网上公开发布 Tableau Desktop 创建的数据分析结果
Tableau Reader	用来打开和阅读其他用户用 Tableau Desktop 创建的数据分析结果

2.5 大数据分析示例——查处虚假出口贸易

本节通过一个示例,说明传统的查询分析与可视化数据分析的主要区别。

2.5.1 案例概述

前面几节我们对大数据分析进行了全面介绍,考虑到本书的侧重点是对大数据的应用分析,为了增强读者的感性认识,我们举一个案例,介绍大数据应用分析的实务流程。

在出口贸易中弄虚作假,以达到走私、骗取出口退税、追求不当收益等目的,是在经济活动中常见的一种违法犯罪行为,也是海关、审计等经济执法和监督机关着力打击的一种犯罪行为。但出口贸易中的违法犯罪行为,手法多样,牵涉的环节多,涉及的数据量大,类型复杂,打击起来难度也很大。例如,有的犯罪分子为了骗取出口退税,向海关虚报出口集装箱,而实际并不出口;有的犯罪分子虚报出口重量,以欺骗手段核销保税料件,以达到掩盖走私保税商品的目的。在执法实践中,有关监管监督机关探索运用大数据分析的方法,打击这类犯罪活动,收到了显著成效。

从进出口货物作业流程来看,完整的通关业务包括海关及以外的诸多单位和部门的共同参与,如商检、外汇管理、税务、商务、海事、空管、港务、码头和银行等,它们和海关一起构成大通关网络。

例如,在查处虚报出口货物重量的过程中,执法人员在对出口流程的调查研究中了解到,海关对企业出口货物的重量监管信息来源于企业申报,如果仅靠这一处数据,缺少印证,无法及时发现不法企业采取多报少出手法虚假出口,骗取出口退税等情况。但是在货物出口通关流程中,所有出口货物在运进出口装货区之前必须在码头的进场卡口处过磅称重,其目的是合理收取费用、保证装货作业安全和运输工具的航行安全,因此码头过磅重量这个外部数据是相对独立、客观、可信的,与海关内部舱单、报关单信息中的出口申报重量数据存在关联关系。这种关联关系在于,扣除出口集装箱的自身箱重和合理误差因素后,实际过磅重量应等于舱单、报关单中的企业申报重量,否则就可能存在不法企业采取多报少出的手法虚假出口、骗取出口退税和走私保税料件的问题。执法人员根据这个分析思路,一环扣一环,进行了步步深入的分析。

本案例分别从海关报关单表、报关单集装箱表、码头出口货物过磅重量表等多处采集企业申报出口货物的重量数据。

本案例选取了其中的部分数据来演示分析过程,对虚报出口货物重量的审计主要用到三张表:海关内部的报关单表、舱单集装箱表和外部码头的码头过磅表。三张表的结构如下:

报关单表(报关单号,进出口标记,进出口日期,航次,出口企业代码,出口企业名称,提单号,报关单申报重量)

舱单集装箱表(进出口标记,航次,提单号,箱号,船号,集装箱申报重量)

码头过磅表(船名,航次,提单号,箱号,申报重量,过磅重量)

把围绕同一个出口申报在不同地方的重量数据进行对比,找出有异常的数据。报关

单与集装箱有三种对应情况：一是一张报关单对应一个集装箱，即一票一箱；二是一张报关单上申报的货物分装在多个集装箱内，形成一张报关单对应多个集装箱的情况，即一票多箱；三是多个企业货物拼装在一个集装箱内，形成多张报关单对应一个集装箱的情况，即拼箱。为简化说明，本案例以一票一箱为例(即一个集装箱仅对应一票提单货物，不包括拼箱和一票多箱)。本案例设定差异下限为 2 000kg，如果超过该值，则将其作为重点关注对象。如果集装箱表申报重量超过其过磅重量的 1.2 倍，则将其视为虚报出口货物重量的疑点重点关注。

2.5.2　查询分析

步骤一：从“舱单集装箱表”(包含全部进出口集装箱基本信息)，提取进出口标记为“出口”(即“进出口标志”＝‘E’)且是一票一箱的集装箱信息，将信息保存在“主表_一票一箱出口舱单表”中。

```
select 船号,航次,提单号
into 主表_一票一箱出口舱单表
from 舱单集装箱表
where 进出口标记='E'      --出口
group by 船号,航次,提单号
having count(*)=1        --一票一箱
```

“主表_一票一箱出口舱单表”包含的部分信息如图 2-5 所示。

	船号	航次	提单号
1	3EZT7	107S	FYCHD-60122
2	3EZT7	107S	NYKS455023441
3	3EZT7	107S	NYKS455024692
4	3EZT7	108S	NYKS455023510
5	3EZT7	108S	NYKS455023545
6	3EZT7	109S	HDFQY-13124
7	3EZT7	109S	HDFQY-13130
8	3EZT7	109S	HDFQY-13133

图 2-5　一票一箱出口舱单表部分数据

```
select * from 主表_一票一箱出口舱单表
```

步骤二：根据已生成的“主表_一票一箱出口舱单表”，结合舱单集装箱表，提取相应提单号所对应的集装箱号集装箱申报重量，将这些信息保存在“主表_一票一箱出口舱单集装箱表”中。

```
select a.船号,a.航次,a.提单号,b.箱号,b.集装箱申报重量
into 主表_一票一箱出口舱单集装箱表
from 主表_一票一箱出口舱单表 a join 舱单集装箱表 b
on a.船号=b.船号 and a.航次=b.航次 and a.提单号=b.提单号
```

“主表_一票一箱出口舱单集装箱表”包含的部分信息如图 2-6 所示。

	船号	航次	提单号	箱号	集装箱申报重量
1	3EZT7	107S	FYCHD-60122	HDMU2183577	9020
2	3EZT7	107S	NYKS455023441	TTNU1305553	20000
3	3EZT7	107S	NYKS455024692	TRLU5458565	11395
4	3EZT7	108S	NYKS455023510	TTNU1305337	23000
5	3EZT7	108S	NYKS455023545	NYKU9353163	4600
6	3EZT7	109S	HDFQY-13124	HDMU2282815	18636
7	3EZT7	109S	HDFQY-13130	HDMU2294035	7380
8	3EZT7	109S	HDFQY-13133	HDMU2231320	11705

图 2-6　一票一箱出口舱单集装箱表部分数据

```
select * from 主表_一票一箱出口舱单集装箱表
```

步骤三：筛选多报少出虚假出口信息，筛选条件：企业申报重量与实际过磅重量之比大于等于120%，企业申报重量与实际过磅重量之差大于2 000kg。将筛选结果保存到“分析表_一票一箱出口集装箱重量异常表”中。

```
select b.船名,a.船号,b.航次,b.提单号,b.箱号 as 集装箱号,
       a.集装箱申报重量,b.过磅重量 * 1000 as 过磅重量,
       a.集装箱申报重量 -b.过磅重量 * 1000 as 多报重量
into 分析表_一票一箱出口集装箱重量异常表
from 主表_一票一箱出口舱单集装箱表 a join 码头过磅表 b
  on a.航次 =b.航次 and a.提单号 =b.提单号 and a.箱号 =b.箱号
where a.集装箱申报重量/b.过磅重量/1000> 1.2
  and a.集装箱申报重量 -b.过磅重量 * 1000> 2000
order by a.集装箱申报重量 -b.过磅重量 * 1000 desc
```

“分析表_一票一箱出口集装箱重量异常表”包含的部分信息如图2-7所示。

```
select * from 分析表_一票一箱出口集装箱重量异常表
```

	船名	船号	航次	提单号	集装箱号	集装箱申报重量	过磅重量	多报重量
1	CSCL ASIA	VRAB8	0001E	XMDFW3AA239	CCLU2519682	17200	13000	4200
2	CSCL ASIA	VRAB8	0001E	XMEMT3AA023	CCLU6370730	17400	14000	3400
3	CSCL ASIA	VRAB8	0001E	XMLAX3AB302	CCLU2419636	16700	13500	3200
4	XIN QIN HUANG DAO	BPBD	0001W	8XMNDUB2A4024	CCLU2281758	17950	14300	3650
5	XIN QIN HUANG DAO	BPBD	0001W	8XMNFXT2A6700	GESU2150149	13339	8700	4639
6	XIN QIN HUANG DAO	BPBD	0001W	8XMNPKG2A5729	CCLU2682597	9600	6200	3400
7	XIN XIA MEN	BPBB	0001W	CXM005387	CAXU7000747	36602	28300	8302
8	XIN XIA MEN	BPBB	0001W	CXM005398	FSCU7679886	25000	19500	5500

图2-7 分析表_一票一箱出口集装箱重量异常表部分数据

步骤四：与“报关单表”进行关联，查找出上述重量异常的出口集装箱对应的报关单记录，并将结果保存到“分析表_一票一箱出口集装箱重量异常报关单”表中。

```
select b.报关单号,a.船名,a.船号,a.航次,a.提单号,b.出口企业名称,
         a.集装箱号,a.多报重量, a.过磅重量,a.集装箱申报重量,b.报关单申报重量
into 分析表_一票一箱出口集装箱重量异常报关单
from 分析表_一票一箱出口集装箱重量异常表 a join 报关单表 b
on a.提单号=b.提单号
order by a.多报重量 desc
```

“分析表_一票一箱出口集装箱重量异常报关单”包含的部分信息如图2-8所示。

```
select * from 分析表_一票一箱出口集装箱重量异常报关单
```

步骤五：数据分析。筛选出多报重量总数最多的5家企业。

```
select top 3 with ties 出口企业名称,sum(多报重量) 多报总重量
from 分析表_一票一箱出口集装箱重量异常报关单
```

	报关单号	船名	船号	航次	提单号	出口企业名称	集装箱号	多报重量	过磅重量	集装箱申报重量	报关单申报重量
1	112004011472O607	OSG ARGOSY	J8JY8	424E	HOXKE424E655	某市7001对外贸易公司	HRSU2301004	5000	20000	25000	24200
2	112004011474 7923	OSG ARGOSY	J8JY8	424E	HOXKE424E681	某地1402有限公司	CLOU2517626	6310	13090	19400	18830
3	112004011495 4334	OSG ARGOSY	J8JY8	425E	HOXKE425E665	某地0238发展公司	SCZU5658212	8000	13500	21500	20425
4	112004011495 4292	OSG ARGOSY	J8JY8	427E	HOXKE427E627	某地0238发展公司	HRSU4301683	5900	15100	21000	19950
5	1120040114773768	OSG ARGOSY	J8JY8	427E	HOXKE427E631	某市0002有限公司	HRSU4303135	3677	15400	19077	18258
6	1120040114777911	OSG ARGOSY	J8JY8	427E	HOXKE427E644	某县1020对外贸易公司	HRSU2302551	13870	13130	27000	26000
7	1120040114956134	OSG ARGOSY	VRAD3	428E	HOXKE428E626	某地0238发展公司	GESU4628622	4100	13500	17600	16800
8	1120040114987606	OSG ARGOSY	VRAD3	432E	HOXKE432E625	某地6676有限公司	TEXU3470183	4500	20500	25000	24500

图 2-8　分析表_一票一箱出口集装箱重量异常报关单部分数据

```
group by 出口企业名称
order by sum(多报重量) desc
```

查询结果如图 2-9 所示。

在对某一个海关的数据分析中,执法人员通过建立上述分析模型进行分析后发现,该关区仅一票一箱的记录中,就有 1 000 多家企业报关出口的 2 000 多个集装箱的货物报关重量与码头过磅数据提供的出口货物称重信息差异较大。经过对该模型运行结果的分析,执法人员选取了关区内的 3 家企业进行延伸检查,以明细出口装箱单为突破口,发现均存在高报单耗、多报出口数量、虚假核销保税料件的问题。相关企业不能提供有效资料证明多核销保税料件的合法去向,涉嫌走私。执法人员依法将其移送当地海关缉私局立案侦查,同时将其余多报少出企业的情况也一并进行了移交,最终对 1 351 家企业进行了处理,严厉打击了违法犯罪行为,维护了正常的贸易秩序。

	出口企业名称	多报总重量
1	某地0012经济贸易公司	356731
2	某市0637有限公司	250690
3	某地8436有限公司	198440
4	某地7022有限公司	197410
5	某地5037发展公司	184792

图 2-9　多报重量总数最多的 5 家企业

2.5.3　可视化分析

本节以 Tableau 10.0 版本作为可视化数据分析平台,介绍在这个环境中分析虚假出口贸易的过程。

步骤一:连接数据源并选择分析的数据。

启动 Tableau 软件,在"连接"窗格,选择"Microsoft SQL Server",在弹出的连接窗口的"服务器"框中输入包含分析数据的 SQL Server 服务器名。

在下一个窗口的"数据库"下拉列表框中选中"大数据_虚假贸易出口"数据库,在列出的表中首先分别双击"报关单表"和"码头过磅表",然后双击"新自定义 SQL",在弹出的窗口中输入如下页所示的 SQL 语句,筛选出一票一箱的集装箱数据(如图 2-10 所示):

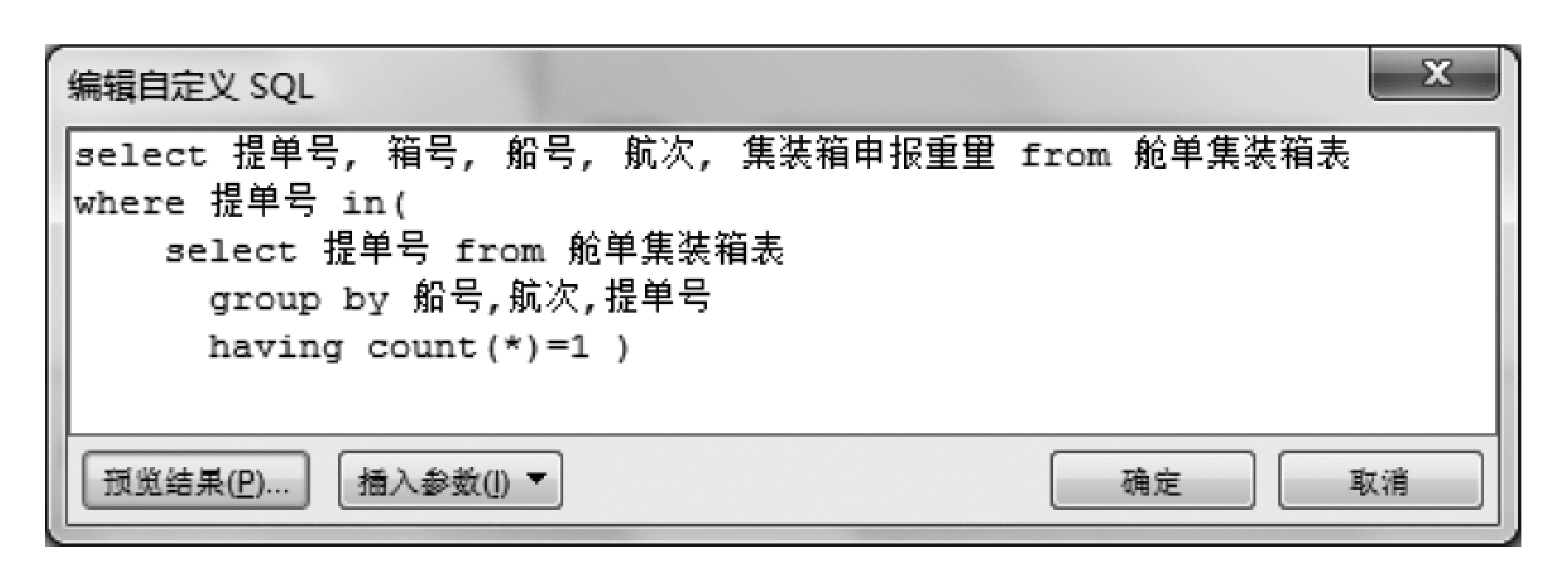


图 2-10　筛选出一票一箱的出口集装箱数据

```
select 提单号,箱号,船号,航次,集装箱申报重量 from 舱单集装箱表
where 提单号 in(
    select 提单号 from 舱单集装箱表
      group by 船号,航次,提单号
      having count(* )=1)
```

在随之出现的“联接”窗口中,指定自定义 SQL 查询产生的数据与已有表之间的关联关系。我们这里在“数据源”下边的下拉列表框中选择“报关单表”中的“提单号”,并在“自定义 SQL 查询”下边的列表框中选中“提单号”,如图 2-11 所示。



图 2-11　指定自定义查询与已有表的关联关系

设置好数据之间连接关系后的“数据源”窗口如图 2-12 所示。



图 2-12　设置好数据之间连接关系后的“数据源”窗口

步骤二:构建计算字段。

在工作表窗格中,为便于分析,创建如下三个计算字段:

(1) 过磅重量(kg)=过磅重量 * 1 000

(2) 多报重量=集装箱申报重量－过磅重量(kg)

（3）申报重量与过磅重量比＝集装箱申报重量/过磅重量(kg)

步骤三：分析数据。

在筛选器中设置筛选条件：

（1）出口标记＝E

（2）多报重量≥2000

（3）申报重量与过磅重量比≥1.2

构建层次结构：船名-航次-提单号

拖放“船名-航次-提单号”到行功能区，并将“出口企业名称”维度拖放到行功能区中“船号-航次-提单号”的右边，将“多报重量”拖放到列功能区中。展示的部分分析结果如图 2-13 所示。

船名	航次 (码头过..	提单号 (码头过磅表)	出口企业名称	多报重量
ACHIM	401N	KMTC-XMN008753	某市0105有限公司	8,800
	403N	AM403NMB673	某地0408有限公司	4,520
		POBUMXB403073904	某地1373发展公司	3,300
	405N	AM405NMB101	某地5022有限公司	16,490
AGIOS DIMITRIOS I	0943W	LTNV854488006511	某地0376有限公司	11,370
	0952W	LTNV854488019737	某县1028有限公司	5,978
ANAN BHUM	S304	XMNCB4040773	某地0491有限公司	2,350
	S305	0264016390	某地0150发展公司	6,020
	S307	XMNCB4041455	某地0459有限公司	5,400
	S308	XMNBKK00622	某市5003进出口公司	7,400
		XMNCB4041752	某地0144有限公司	2,600
		XMNCB4041825	某地6004有限公司	5,762
		XMNHCN00600	某市0227有限责任公司	3,255
ANL AUSTRALIA	021S	ANLU865921160	某市0676有限公司	2,360
		CXMNBNE2A4115	某地0876有限公司	10,710
		OOLU81128030	某市0002对外贸易公司	2,629
	022S	8XMNMEL2A4359	某市4877有限公司	2,440
		OOLU90469630	某县6014有限公司	10,848
ANL ESPRIT	002S	8XMNSYD2A4367	某市4241有限公司	5,550
ANL EXPLORER	021S	ANLU865921049	某县6014有限公司	4,184
		CXMNSYD2A4223	某地0491有限公司	3,387
	022S	8XMNBNE2A4132	某市0676有限公司	2,250

0K 20K 40K 60K 80K 100K 120K 140K 160K 180K
多报重量

图 2-13　展示多报重量的企业及多报重量

图 2-14 展示了按多报重量数降序排序的各出口企业名称及其多报重量合计。

图 2-15 用条形图的形式展示了不同多报重量数据段的出口企业个数。这里将多报重量划分为 10 个组，对多报重量少于 10 000 的，以 1 000kg 为递增值划分数据段，对多报重量超过 10 000 的，将多报重量划分为 10 001～12 000 及大于 12 000 两个数据段。条形图上显示的数据代表在这个数据段内多报重量的企业个数。

Tableau 提供了多种数据展示形式，图 2-16 为图 2-15 的折线图展示形式。折线图有助于用户发现数据的变化趋势。

图 2-17 为图 2-15 的气泡图展示形式。气泡图用圆圈的不同大小来揭示数据的意义。

2.5.4　分析小结

从前边查询分析和可视化分析的分析过程，可看到传统的分析方法与大数据环境下的可视化分析方法有如下几个主要区别(这里仅以结构化数据为例进行对比)。

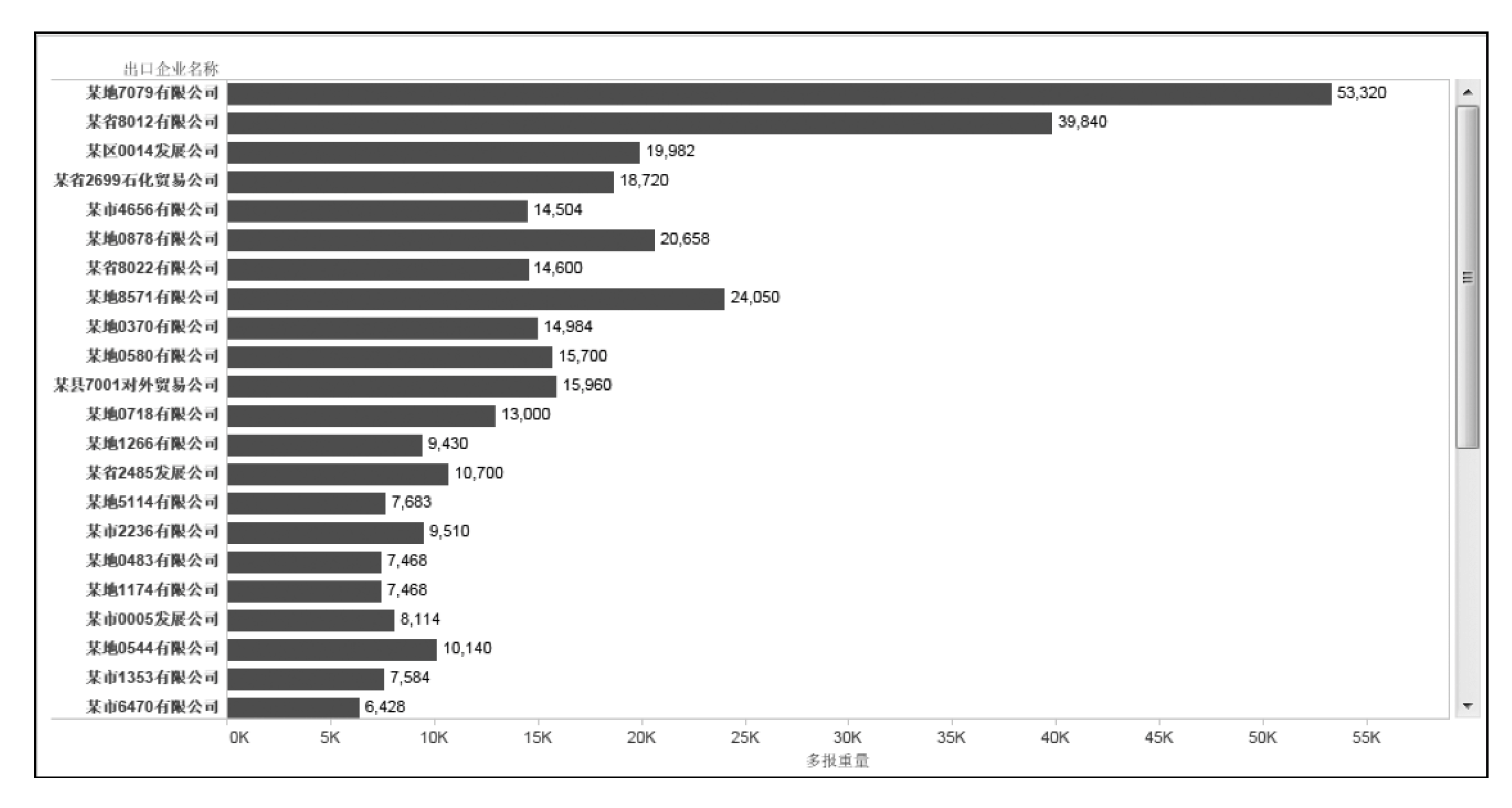


图 2-14　按多报重量降序展示企业及其多报的总重量

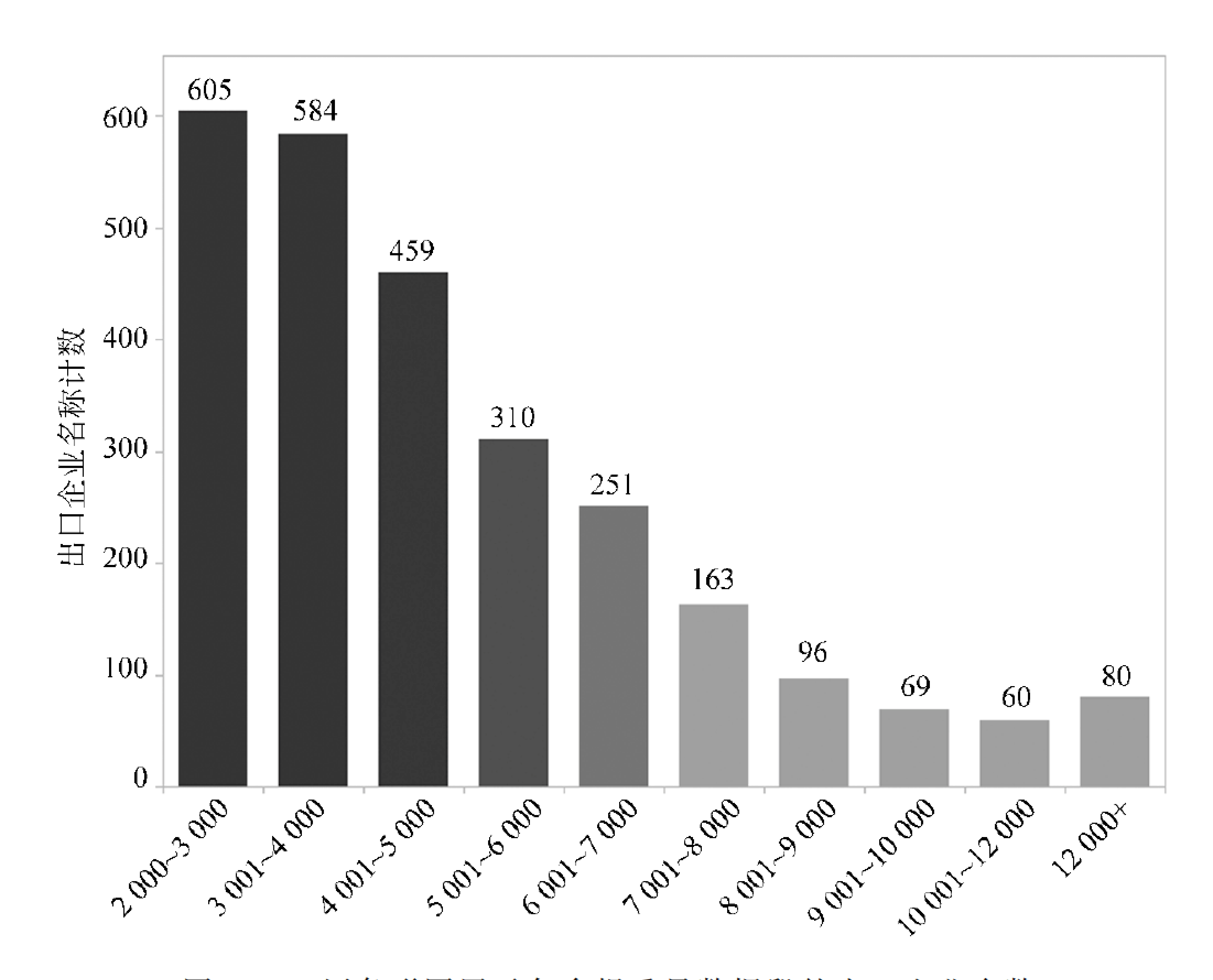


图 2-15　用条形图展示各多报重量数据段的出口企业个数

1. 对分析人员的要求不同

- 传统的查询分析要求分析人员必须具有扎实良好的数据库知识，特别是要具有比较好的编写查询语句的能力，能够熟练构建大量分析模型，否则数据分析将无从下手。
- 可视化分析分析则主要要求分析人员进行拖拖放放的简单操作，即可构建数据的分析模型，编写查询语句、构建分析模型的要求相对简单。

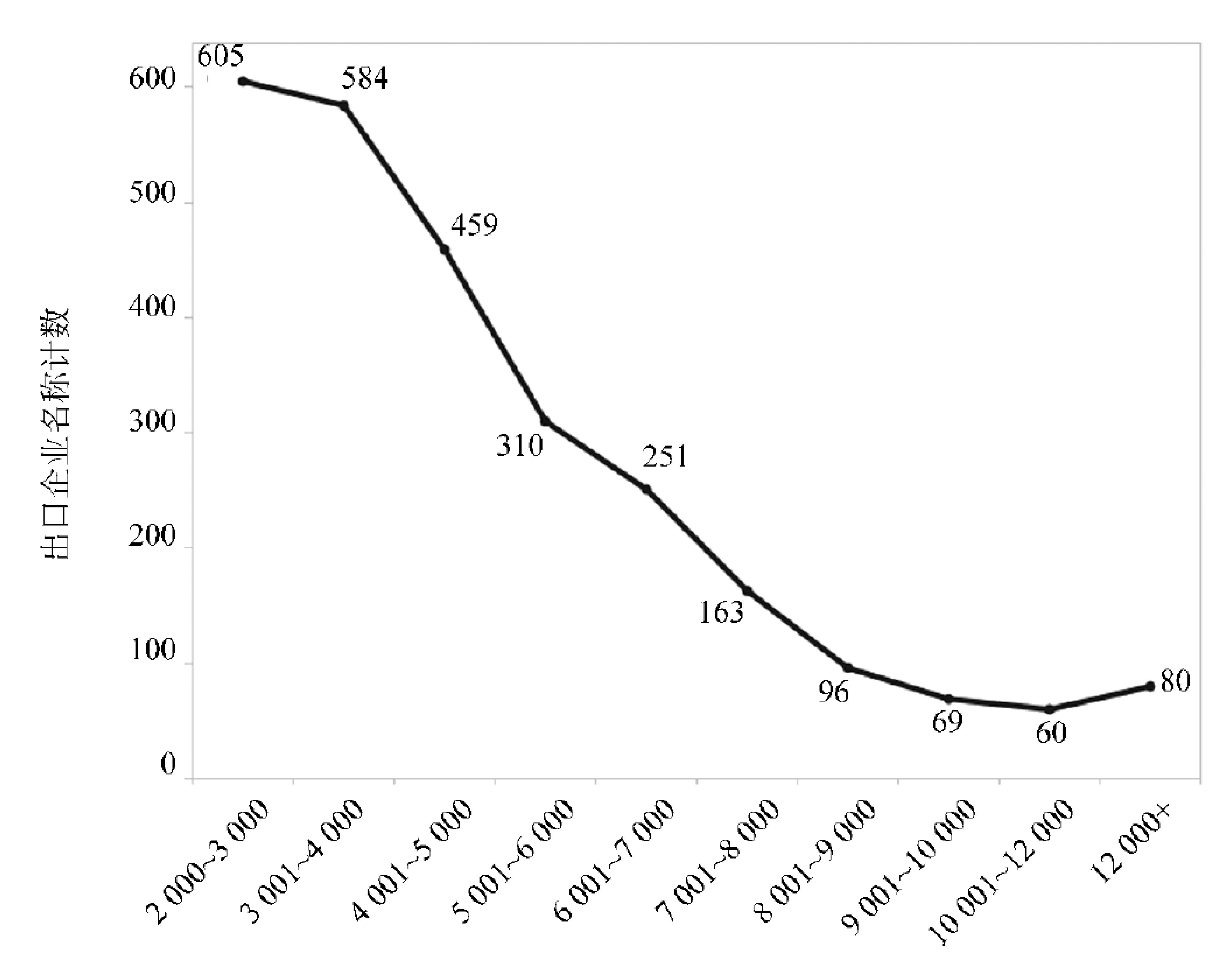


图 2-16　用折线图展示各多报重量数据段的出口企业个数

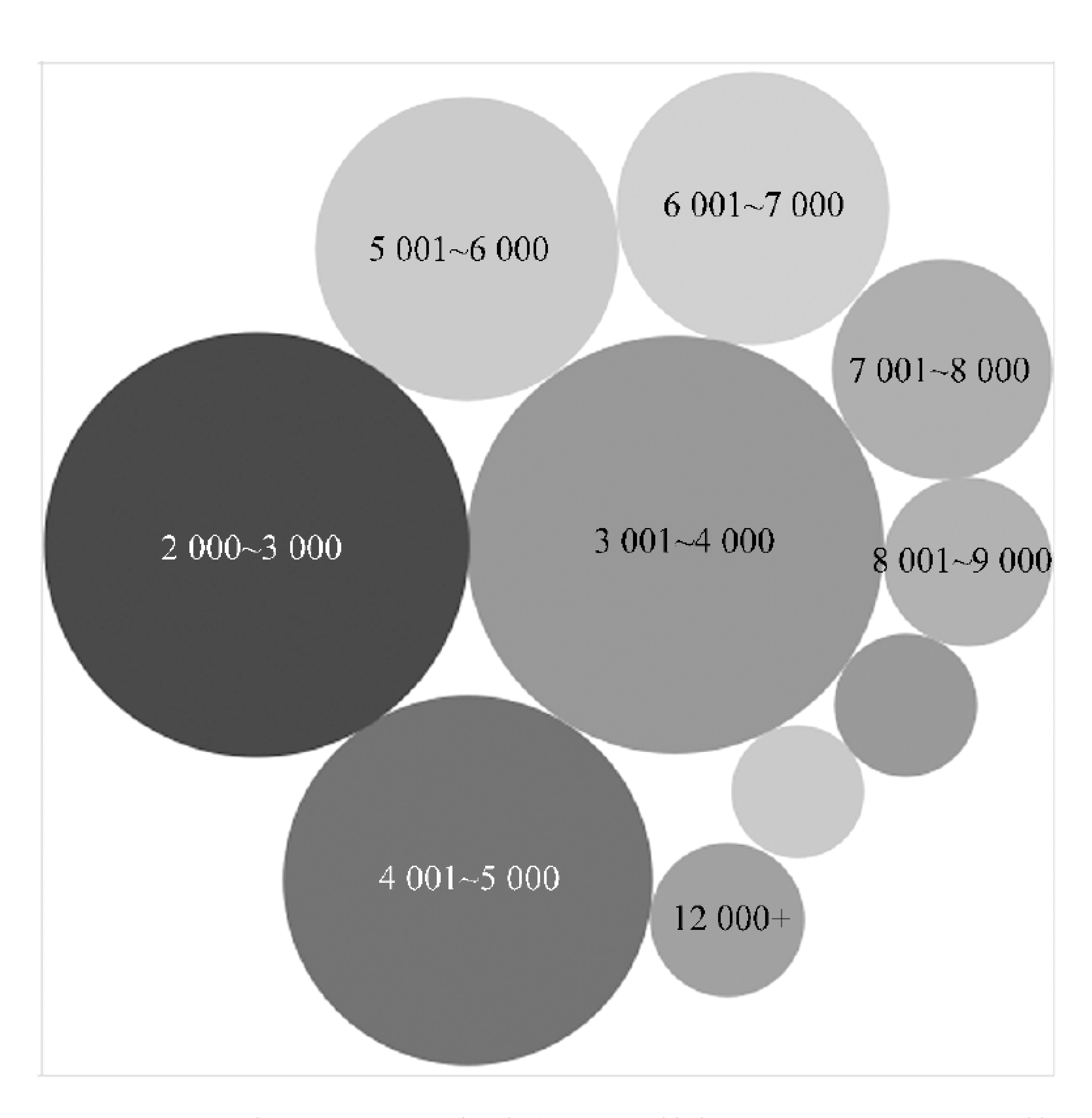


图 2-17　用气泡图展示各多报重量数据段的出口企业个数

2. 分析结果的展示不同

- 传统的查询分析的展示结果只能是二维表的形式，分析结果不直观。
- 可视化分析的展示结果可以是花样繁多的各种图的形式，分析结果直观，易于理解、对比和了解趋势。

3. 数据实时性不同

- 传统的查询分析在产生查询结果后，其结果就与源数据没有关系了。源数据发生变化，如果不重新执行查询语句，则分析结果不会改变。因此，传统的查询分析方法的分析结果不能实时反映数据的变化情况，其分析结果与实时数据有一定的滞后。
- 可视化数据分析方法的分析结果是可以随源数据的变化而实时自动更新的，因此大数据环境下的可视化数据分析展示的可以是实时数据的分析结果，这种实时性不需要用户做任何工作，是由可视化分析软件自动实现的。

参考文献

[1] 李学龙，龚海刚. 大数据系统综述[J]. 中国科学：信息科学，2015，(1)：1-44.
[2] [英]迈尔—舍恩伯格，库克耶. 大数据时代[M]. 盛杨燕，周涛，译. 杭州：浙江人民出版社，2013.
[3] 刘鹏. 云计算：第 3 版[M]. 北京：电子工业出版社，2015.
[4] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，(1)：146-169.
[5] 刘智慧，张泉灵. 大数据技术研究综述[J]. 浙江大学学报：工学版，2014，(6)：957-972.
[6] 梁吉业，冯晨娇，宋鹏. 大数据相关分析综述[J]. 计算机学报，2016，(1)：1-18.

第 3 章　常用数据分析与预测方法

3.1　方差分析

3.1.1　分析方法

方差分析(Analysis of Variance,ANOVA),又称变异数分析或 F 检验,是英国统计学家罗纳德·艾尔默·费希尔(R. A. Fisher)于 1923 年发明的统计方法。方差分析研究诸多因素中哪些因素对观测变量有显著影响,在科学试验和现代化工业质量控制中得到了广泛的应用。

一个复杂的事物,其中往往有许多因素互相制约又互相依存。方差分析的目的是通过数据分析找出对该事物有显著影响的因素、各因素之间的交互作用以及显著影响因素的最佳水平等。常用的方差分析方法包括单因素方差分析、多因素方差分析、多元方差分析、协方差分析、重复设计方差分析。单因素方差分析是研究一个因素的变化是否对事物产生了显著的影响。在实际的应用中,一个事物往往受多个因素的影响。多因素方差分析是对一个独立因素是否受一个或多个其他因素影响而进行的方差分析。多因素方差分析不仅能分析每个因素对事物的影响,还能分析各个因素间的交互作用对事物是否有显著的影响。例如,应用多因素方差对大棚作物产量进行分析时,光照、湿度、温度对作物的产量都会有很大的影响,而光照、湿度、温度的交互作用对作物最终的产量的影响更显著。

本节介绍的示例是应用方差分析法分析职业和性别对薪资的影响,这属于多因素方差分析中的双因素方差分析。根据双因素方差分析中两个因素是否相互影响,将其分为可重复和无重复的双因素分析。无重复的双因素分析表示两个因素对结果的影响是相互独立的;可重复的双因素分析表示两个因素除了对结果单独影响外,二者的搭配还会对结果产生新的影响。本示例按照可重复的双因素方差分析进行处理。

本示例采用 Excel 2013 进行分析,分析时需要考虑两个参数:相伴概率 p 与显著性水平 α。对于某个因数 A,若 $p<\alpha$ 则因数 A 对变量有显著性影响,反之则影响不显著。对于因数 A 与 B 的交互作用,若 $p<\alpha$ 则因数 A 与 B 的交互作用对变量有显著性影响,反之则没有显著性影响。

3.1.2　示例介绍

某杂志的记者想考察职业为财务管理、计算机程序员和药剂师的男、女雇员每周的薪资是否有显著性的差异。从每种职业中分别选取了 5 名男性和 5 名女性组成样本,并且记录下样本中每个人的周薪资,周薪资的单位是美元,样本数据如表 3-1 所示。现要分析职业和性别对薪资有无显著性影响。

表 3-1　薪资数据

职　　业	性　　别	每周薪资
财务管理	男	872
财务管理	男	859
财务管理	男	1 028
财务管理	男	1 117
财务管理	男	1 019
财务管理	女	519
财务管理	女	702
财务管理	女	805
财务管理	女	558
财务管理	女	591
计算机程序员	男	747
计算机程序员	男	766
计算机程序员	男	901
计算机程序员	男	690
计算机程序员	男	881
计算机程序员	女	884
计算机程序员	女	765
计算机程序员	女	685
计算机程序员	女	700
计算机程序员	女	671
药剂师	男	1 105
药剂师	男	1 144
药剂师	男	1 085
药剂师	男	903
药剂师	男	998
药剂师	女	813
药剂师	女	985
药剂师	女	1 006
药剂师	女	1 034
药剂师	女	817

3.1.3　示例分析

在 Excel 2013 中，对该数据进行如下分析。

(1) 将表 3.1 中的数据录入 Excel 文件中，该文件中的数据样式如图 3-1 所示。

	A	B	C	D
1		男	女	
2	财务管理	872	519	
3		859	702	
4		1028	805	
5		1117	558	
6		1019	591	
7	计算机程序员	747	884	
8		766	765	
9		901	685	
10		690	700	
11		881	671	
12	药剂师	1105	813	
13		1144	985	
14		1085	1006	
15		903	1034	
16		998	817	
17				

图 3-1　“方差分析_薪资表”文件中的数据

(2) 在包含该数据的 Excel 文件中，单击“数据”功能区中最右边的“数据分析”。在弹出的“数据分析”窗口中选择“方差分析：可重复双因素分析”(如图 3-2 所示)，单击“确定”按钮，弹出如图 3-3 所示的设置参数窗口。

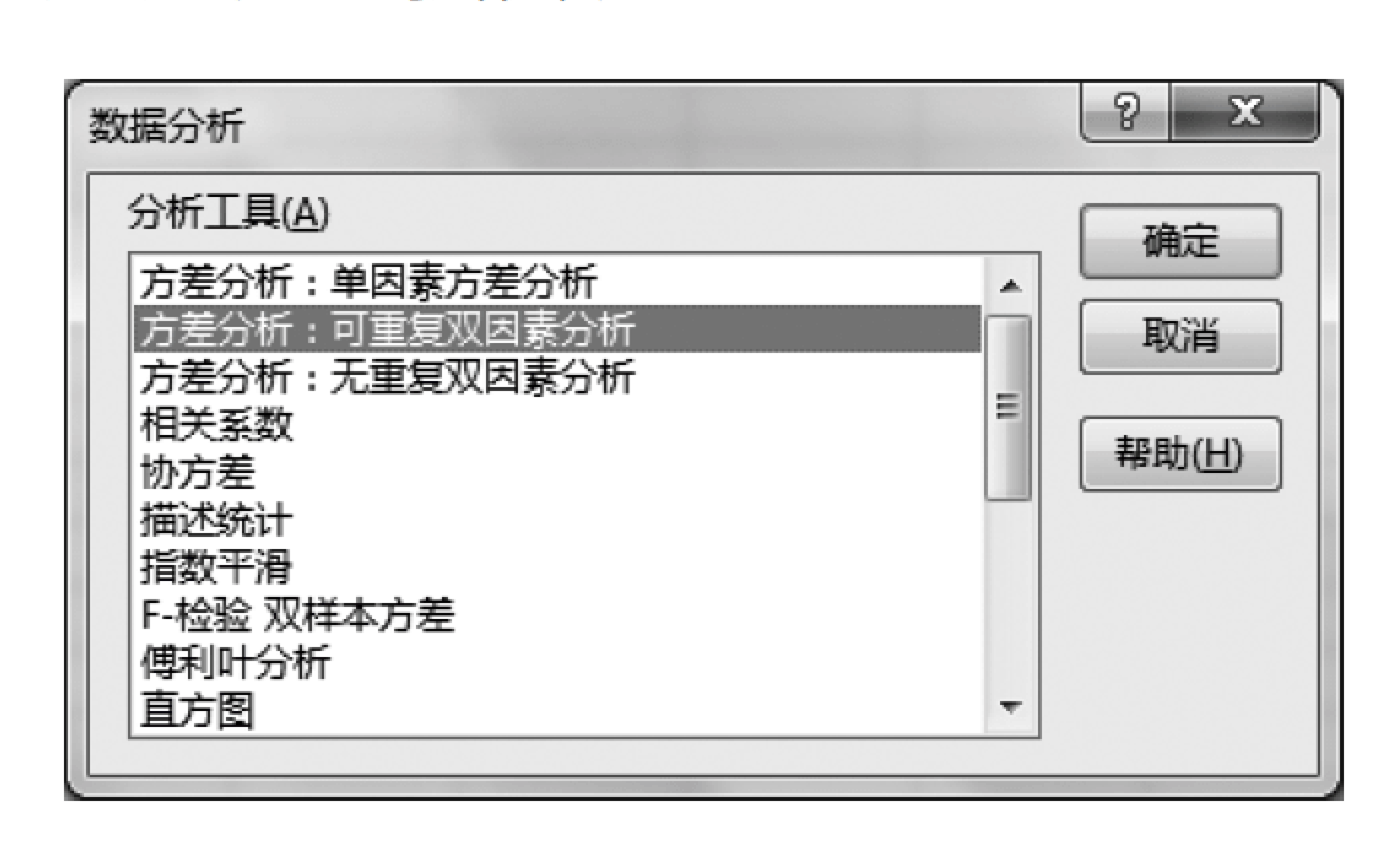


图 3-2　选择“方差分析：可重复双因素分析”

(3) 在如图 3-2 所示的窗口配置相关系数。在输入区域输入：A1:C16；每一样本的行数中输入：5；显著性水平 α(A)中输入：0.05；在“输出选项”中，选“输出区域新工作表”，并输入“A18”，我们从 A18 单元格开始显示方差分析结果。设置好后的情形如图 3-3 所示。

(4) 设置好参数后单击“确定”按钮，表格中将显示分析结果。如图 3-4 所示为方差分析结果，如图 3-5 所示为分析的汇总结果。

(5) 计算职业对薪资的显著性影响。

选中单元格 A55，在公式栏中输入如下代码(如图 3-6 所示)：

```
=IF(F48<0.05,"职业对薪资有差异","职业对薪资无差异")
```

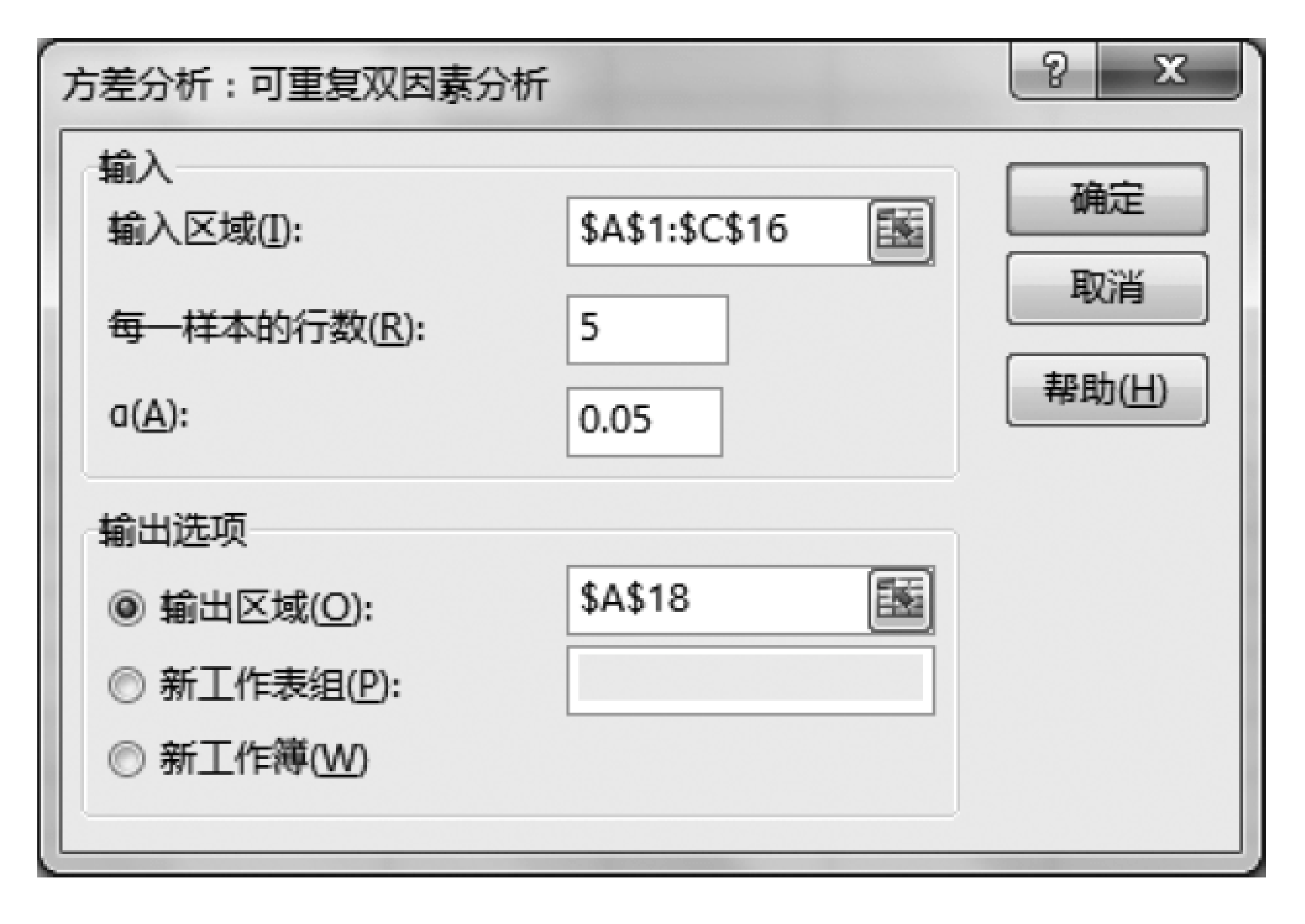


图 3-3　配置相关系数

	A	B	C	D	E	F	G
46	方差分析						
47	差异源	SS	df	MS	F	P-value	F crit
48	样本	276560	2	138280	13.24563	0.000133	3.402826
49	列	221880	1	221880	21.25355	0.000112	4.259677
50	交互	115440	2	57720	5.528912	0.010595	3.402826
51	内部	250552	24	10439.67			
52							
53	总计	864432	29				

图 3-4　方差分析结果

	A	B	C	D	E
18	方差分析：可重复双因素分析				
19					
20	SUMMARY	男	女	总计	
21	财务管理				
22	观测数	5	5	10	
23	求和	4895	3175	8070	
24	平均	979	635	807	
25	方差	12223.5	13677.5	44382.67	
26					
27	计算机程序员				
28	观测数	5	5	10	
29	求和	3985	3705	7690	
30	平均	797	741	769	
31	方差	8195.5	7685.5	7929.333	
32					
33	药剂师				
34	观测数	5	5	10	
35	求和	5235	4655	9890	
36	平均	1047	931	989	
37	方差	9338.5	11517.5	13007.11	
38					
39	总计				
40	观测数	15	15		
41	求和	14115	11535		
42	平均	941	769		
43	方差	20436.42857	25460.14286		
44					

图 3-5　汇总结果

A55	fx	=IF(F48<0.05,"职业对薪资有差异","职业对薪资无差异")					
	A	B	C	D	E	F	G

图 3-6　职业对薪资显著性影响计算公式

输完代码后按回车键，职业对薪资的显著性影响结果将显示在 A55 单元格中。
(6) 计算性别对薪资的显著性影响。
选中单元格 A56，在公式栏中输入如下代码：

```
=IF(F49<0.05,"性别对薪资有差异","性别对薪资无差异")
```

输完代码后按回车键，职业对薪资的显著性结果出现在 A56 单元格中。
(7) 计算职业与性别的交互作用。
选中单元格 A57，在公式栏中输入如下代码：

```
=IF(F50<0.05,"职业与性别有交互作用","职业与性别无交互作用")
```

输完代码后按回车键，职业对薪资的显著性结果出现在 A57 单元格中。
步骤(5)、(6)、(7)的校验结果如图 3-7 所示。

	A	B	C	D	E	F	G
46	方差分析						
47	差异源	SS	df	MS	F	P-value	F crit
48	样本	276560	2	138280	13.24563	0.000133	3.402826
49	列	221880	1	221880	21.25355	0.000112	4.259677
50	交互	115440	2	57720	5.528912	0.010595	3.402826
51	内部	250552	24	10439.67			
52							
53	总计	864432	29				
54							
55	职业对薪资有差异						
56	性别对薪资有差异						
57	职业与性别有交互作用						

图 3-7　校验结果

3.1.4　结果分析与总结

图 3-7 反映的在显著性水平 $\alpha=0.05$ 的条件下，职业、性别以及职业与性别的交互作用是否对每周的薪资产生了显著性影响。

对于职业对薪资的影响，由于 $p=0.000\ 133<0.05$，所以职业对薪资有显著性的影响。

对于性别对薪资的影响，由于 $p=0.000\ 112<0.05$，所以性别对薪资有显著性的影响。

对于职业和性别二者的交互作用对薪资的影响，由于 $p=0.010\ 595<0.05$，所以职业和性别的交互作用对薪资也产生了显著性的影响。

因此本示例中，职业、性别以及职业与性别的交互作用对每周的薪资都有显著性的影响。

3.2　相关分析

3.2.1　分析方法

事物之间往往存在某种关联性，如果这种关联性可以用函数表示，则称它们之间是一种函数关系。现实中很多事物之间虽然存在某种联系，但不能应用已知的函数关系来表示，这种联系即为相关关系。如果这种相关性只涉及两个事物则为单相关，如果涉及三个

或者三个以上的事物则为复相关、多重相关。

事物之间的相关程度使用相关系数来衡量，相关系数表示事物之间关系的紧密程度。对于复相关，往往采用多重相关系数考察一个变量与其他变量之间的相关程度，采用偏相关系数考察多个变量中两个变量的相关性。

在有 n 个($n\geqslant 3$)变量的系统中，若要考察第 i 个变量与其余 $n-1$ 个变量的相关程度，采用多重相关系数来表示，计算公式为$\sqrt{1-\frac{R}{R_{ii}}}$。$R$ 是单相关系数矩阵对应的行列式，R_{ii}是R 的第i 行、第 i 列的代数余子式。R_{ii}的代数余子式是在去掉R 中的第i 行与第 i 列元素后得到的行列式。同理求 R_{ij} 的代数余子式是通过去掉R 中的第i 行与第j 列元素得到的行列式。

多重相关性中考察一个变量与另外一个变量之间的相关性用偏相关系数来表示。例如，考察变量 i 与变量 j 之间的偏相关性，计算公式为$(-1)^{(i+j)}\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}}$。该值的绝对值越大表明变量 i 和 j 的偏相关程度越大，二者的关系越紧密，相互影响越明显。

本相关分析方法将用包含股票价格、成交金额、收益率三个变量的示例来说明。本示例也采用 Excel 中的数据分析工具对三个变量的相关性进行分析。

3.2.2 示例介绍

某上市公司 8 月前 15 个交易日的收益率、股票价格和成交金额样本数据如表 3-2 所示。现要计算：

(1) 收益率与股票价格和成交金额的多元相关系数；

(2) 收益率与股票价格的偏相关系数；

(3) 收益率与成交金额的偏相关系数。

表 3-2 8 月份股票交易样本数据

日期	股票价格/元	成交金额/元	收益率
20080801	9.28	41 652 766	0.014 923
20080802	9.23	18 716 130	−0.002 3
20080803	9.18	41 314 097	−0.023 15
20080804	8.96	18 393 783	0.003 234
20080805	8.95	34 259 522	−0.007 53
20080806	8.73	31 981 311	−0.025 97
20080807	8.65	43 000 708	−0.011 11
20080808	8.59	35 314 780	0.011 236
20080809	8.52	34 774 469	0.039 535
20080810	8.49	32 888 399	0.002 237
20080811	8.42	23 306 213	−0.018 97

续表

日　　期	股票价格/元	成交金额/元	收益率
20080812	8.37	38 787 086	−0.014 79
20080813	8.31	30 253 320	0.020 785
20080814	8.26	41 662 276	−0.002 26
20080815	8.21	23 703 595	0.002 245

3.2.3　示例分析

在 Excel 2013 中，对该数据进行如下分析。

（1）将如表 3-2 所示的数据录入 Excel 文件中，该文件中的数据样式如图 3-8 所示。

	A	B	C	D
1	股票价格（元）	成交金额（元）	收益率	
2	9.28	41652766	0.014923	
3	9.23	18716130	-0.0023	
4	9.18	41314097	-0.02315	
5	8.96	18393783	0.003234	
6	8.95	34259522	-0.00753	
7	8.73	31981311	-0.02597	
8	8.65	43000708	-0.01111	
9	8.59	35314780	0.011236	
10	8.52	34774469	0.039535	
11	8.49	32888399	0.002237	
12	8.42	23306213	-0.01897	
13	8.37	38787086	-0.01479	
14	8.31	30253320	0.020785	
15	8.26	41662276	-0.00226	
16	8.21	23703595	0.002245	
17				

图 3-8　股票交易数据

（1）在该文件中，单击“数据”，然后单击“数据”功能区最右边的“数据分析”图标。在弹出的“数据分析”窗口中选择“相关系数”（如图 3-9 所示），单击“确定”按钮，弹出如图 3-10 所示的设置参数窗口。



图 3-9　选择“相关系数”

（2）在如图 3-10 所示的“相关系数”窗口中，在“输入区域”输入：A1：C16；在“分组方式”中选中“逐列”；勾选“标志位于第一行”；在“输出选项”部分选中“输出区域”，并在后边的文本框中输入：A18。

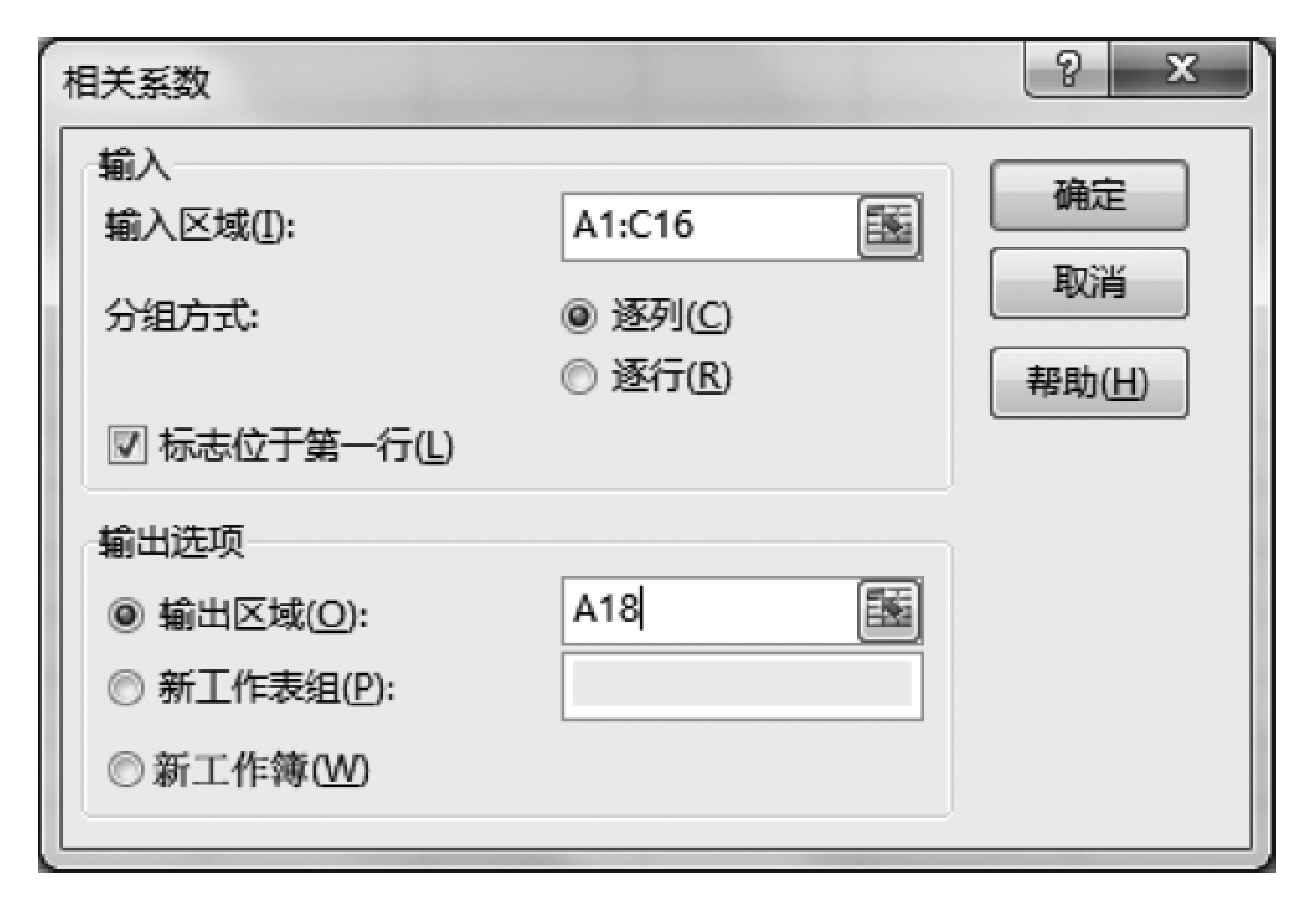


图 3-10　配置相关系数

（3）单击“确定”按钮，从单元格 A18 开始将显示如图 3-11 所示的矩阵信息。

18		股票价格（元）	成交金额（元）	收益率
19	股票价格（元）	1		
20	成交金额（元）	-0.01476237	1	
21	收益率	-0.135232122	-0.038651537	1

图 3-11　单相关系数矩阵

（4）根据对称性填充图 3-11 中矩阵上方的空单元格，结果如图 3-12 所示。

18		股票价格（元）	成交金额（元）	收益率
19	股票价格（元）	1	-0.01476237	-0.135232122
20	成交金额（元）	-0.01476237	1	-0.038651537
21	收益率	-0.135232122	-0.038651537	1

图 3-12　矩阵填充结果

（5）列出 R 的行列式以及 R_{11}、R_{22}、R_{33}、R_{13}、R_{23} 的代数余子式，如图 3-13 所示。

	A	B	C	D
24		单相关系数矩阵		
25	1	-0.01476237	-0.135232122	
26	-0.01476237	1	-0.038651537	
27	-0.135232122	-0.038651537	1	
28				
29	R_{11}			
30	1	-0.038651537		
31	-0.038651537	1		
32				
33	R_{22}			
34	1	-0.135232122		
35	-0.135232122	1		
36				
37	R_{33}			
38	1	-0.01476237		
39	-0.01476237	1		
40				
41	R_{13}			
42	-0.01476237	1		
43	-0.135232122	-0.038651537		
44				
45	R_{23}			
46	1	-0.01476237		
47	-0.135232122	-0.038651537		

图 3-13　各个矩阵数据

(6) 计算 R 的行列式的值。

选中单元格 B49，然后在公式栏中输入如下公式(如图 3-14 所示)：

```
=MDETERM(A25:C27)
```

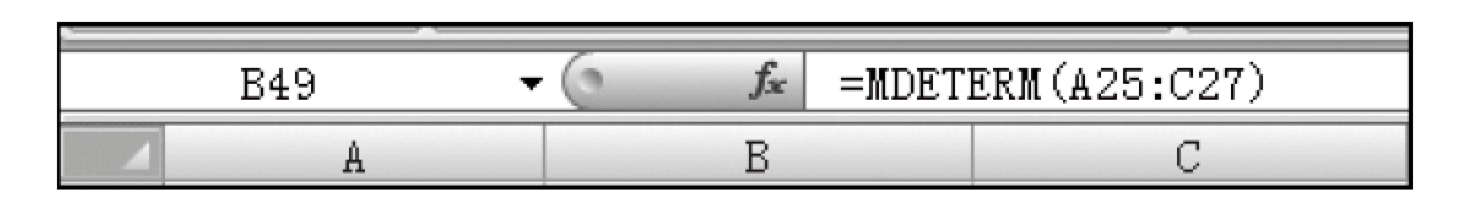


图 3-14　计算 R 的公式

输完代码后按回车键，计算结果如图 3-15 所示。

(7) 计算 R_{11} 的行列式的值。

选中单元格 B50，在公式栏中输入如下公式并按回车键，计算结果如图 3-13 所示。

```
=MDETERM(A30:B31)
```

输完代码后按回车键，计算结果如图 3-15 所示。

(8) 计算 R_{22} 的行列式的值。

选中单元格 B51，并在公式栏中输入如下公式，并按回车键，计算结果如图 3-15 所示。

```
=MDETERM(A34:B35)
```

(9) 计算 R_{33} 的行列式的值。

选中单元格 B52，在公式栏中输入如下公式，并按回车键，计算结果如图 3-15 所示。

```
=MDETERM(A38:B39)
```

(10) 计算 R_{13} 的行列式的值。

选中单元格 B53，在公式栏中输入如下公式，并按回车键，计算结果如图 3-15 所示。

```
=MDETERM(A42:B43)
```

(11) 计算 R_{23} 的行列式的值。

选中单元格 B54，在公式栏中输入如下公式，并按回车键，计算结果如图 3-15 所示。

```
=MDETERM(A46:B47)
```

	A	B	C
49	R	0.979846081	
50	R_{11}	0.998506059	
51	R_{22}	0.981712273	
52	R_{33}	0.999782072	
53	R_{13}	0.13580271	
54	R_{23}	-0.040647883	
55			

图 3-15　R、R_{11}、R_{22}、R_{33}、R_{13}、R_{23} 的值

(12) 计算收益率与股票价格和成交金额的多元相关系数：$\sqrt{1-\frac{R}{R_{33}}}$。

选中单元格 B56,在公式栏中输入如下公式,并按回车键,计算结果如图 3-16 所示。

```
=SQRT(1-B49/B52)
```

(13) 计算收益率与股票价格的偏相关系数:$(-1)^{(1+3)}\dfrac{R_{13}}{\sqrt{R_{11}R_{33}}}$。

选中单元格 B57,在公式栏中输入如下公式,并按回车键,计算结果如图 3-16 所示。

```
=(-1)^(1+3)*B53/SQRT(B50*B52)
```

(14) 计算收益率与成交金额的偏相关系数:$(-1)^{(2+3)}\dfrac{R_{23}}{\sqrt{R_{22}R_{33}}}$。

选中单元格 B58,在公式栏中输入如下公式,并按回车键,计算结果如图 3-16 所示。

```
=(-1)^(2+3)*B54/SQRT(B51*B52)
```

	A	B	C
56	多重相关系数	0.141210259	
57	偏相关系数1	0.135919076	
58	偏相关系数2	0.04102921	

图 3-16　多重相关和偏相关系数值

3.2.4　结果分析与总结

图 3-16 反映的是收益率、股票价格和成交金额三者之间的相关性。单元格 B56 中反映的是收益率与股票价格和成交金额的多重相关系数,该值越大表明收益率与股票价格和成交金额的线性相关程度越密切。单元格 B57 与 B58 分别是收益率与股票价格、收益率与成交金额的偏相关系数值。偏相关系数用于多要素组成的系统中,单独考察一个要素对其他要素的影响。其值取值范围介于-1 和 1 之间,绝对值越大表明其偏相关的程度越大。本示例中,相较于成交金额,股票价格对收益率的影响更大。

3.3　回归分析

3.3.1　分析方法

回归分析是一种统计分析方法,用于确定变量之间的函数关系,主要用于数据的预测。回归分析方法的思想是根据若干个变量的一系列的实际观测值,推断出这些变量之间存在的函数关系,然后再利用所获得的函数关系预测某个变量的取值。如果回归分析只涉及两个变量且二者的关系可以表示为线性函数时则称之为一元线性回归分析;如果回归分析中包含三个或三个以上的变量且变量之间可以表示成线性函数则称之为多重线性回归分析。

进行回归分析时,需要使用残差来衡量回归分析结果的优劣。残差是预测值和实际观测值之间的差额。当我们获得了一个回归分析的函数关系时,对于给定的自变量,可以计算出因变量的值。但是这种函数只是尝试去逼近真实的情况,由于随机误差等因素,根

据函数关系计算得到的因变量的值(又称预测值)与实际观测值有一定的差距。残差就是用来衡量其大小的指标。残差越小,说明预测值和实际观测值越接近,回归分析的结果也越好。

本节给出的示例是考察公司收入与电视和报纸的广告费用间的关系,属于多重线性回归分析的范畴。

3.3.2　示例介绍

某媒体公司的管理者认为公司每周的收入与广告费用是密切相关的,他们想对每周的总收入做出预测和评估。这家公司收集获得了 8 周的历史数据组成样本数据,如表 3-3 所示。

表 3-3　收入与电视广告费、报纸广告费关系数据　　单位:千元

每周的总收入	电视广告费用	报纸广告费用
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.5
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

现要进行如下两项工作:

(1) 试通过表中的数据给出广告费用与收入的回归方程;

(2) 在显著水平为 0.05 时,对方程进行总体显著性和回归系数的显著性检验。

3.3.3　示例分析

在 Excel 2013 中,对该数据进行如下分析。

(1) 将如表 3-3 所示的数据录入 Excel 文件中,该文件中的数据样式如图 3-17 所示。

	A	B	C	D
1	样本数	每周的总收入/千元	电视广告费用/千元	报纸广告费用/千元
2	1	96	5	1.5
3	2	90	2	2
4	3	95	4	1.5
5	4	92	2.5	2.5
6	5	95	3	3.5
7	6	94	3.5	2.3
8	7	94	2.5	4.2
9	8	94	3	2.5
10				

图 3-17　收入与电视、报纸广告费用数据

（2）在该文件的“数据”功能区中，单击右边的“数据分析”图标。在弹出的“数据分析”窗口中选择“回归”（如图 3-18 所示），单击“确定”按钮，弹出如图 3-19 所示的设置参数窗口。



图 3-18　选择“回归”

（3）在弹出的“回归”框中配置相关系数。在 Y 值输入区域输入：B1:B9；X 值输入区域输入：C1：D9；勾选“标志”和“置信度”，“置信度”中输入 95；在“输出区域”的文本框中输入：A12；在“残差”部分勾选“残差”“残差图”“标准残差”和“线性拟合图”。配置好相关系数的回归窗口如图 3-19 所示。



图 3-19　配置相关系数

（4）单击“确定”按钮后，回归分析结果便出现在 A12 开始的下方单元格中。回归汇总分析结果如图 3-20 所示，残差输出如图 3-21 所示，电视广告费用残差图与线性拟合图如图 3-22 所示，报纸广告费用残差图与线性拟合图如图 3-23 所示。

3.3.4　结果分析与总结

根据图 3-20 中数据表“Coefficients”列的相关数据，可以得到回归方程为 $Y=83.28+2.28X_1+1.27X_2$，其中 Y 表示收入，X_1 表示电视广告费用，X_2 表示报纸广告费用。

当回归的显著性水平为 0.05 时，方程总体拟合优度为 0.90，且通过 F 检验，因此回归方程总体显著。

	A	B	C	D	E	F	G	H	I
12	SUMMARY OUTPUT								
13									
14	回归统计								
15	Multiple R	0.963955986							
16	R Square	0.929211142							
17	Adjusted R Square	0.900895599							
18	标准误差	0.600852041							
19	观测值	8							
20									
21	方差分析								
22		df	SS	MS	F	gnificance F			
23	回归分析	2	23.69488412	11.84744206	32.81629	0.001333			
24	残差	5	1.805115875	0.361023175					
25	总计	7	25.5						
26									
27		Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
28	Intercept	83.28284245	1.442065966	57.75244991	2.94E-08	79.57589	86.98979	79.57589	86.98979
29	电视广告费用/千元	2.283844253	0.281907565	8.101393999	0.000465	1.559178	3.008511	1.559178	3.008511
30	报纸广告费用/千元	1.274961598	0.288418209	4.420530872	0.006889	0.533559	2.016364	0.533559	2.016364

图 3-20　回归汇总分析结果

	A	B	C	D
34	RESIDUAL OUTPUT			
35				
36	观测值	测 每周的总收入/千	残差	标准残差
37	1	96.61450611	-0.614506111	-1.210103963
38	2	90.40045415	-0.400454151	-0.788586389
39	3	94.33066186	0.669338142	1.318080852
40	4	92.17985708	-0.179857076	-0.354179977
41	5	94.5967408	0.4032592	0.794110176
42	6	94.20870901	-0.20870901	-0.410996075
43	7	94.34729179	-0.347291792	-0.683897469
44	8	93.3217792	0.678220797	1.335572845

图 3-21　残差输出结果

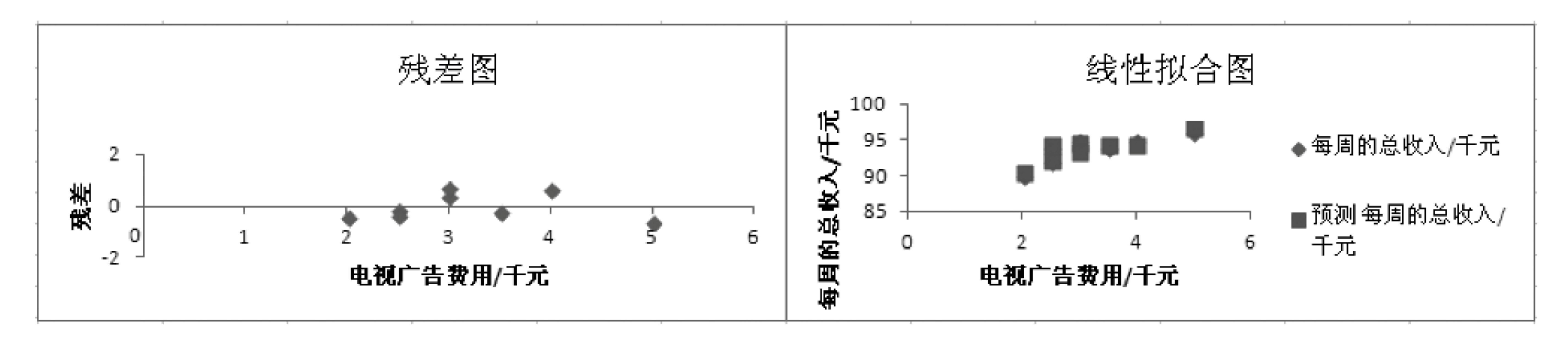


图 3-22　电视广告费用残差图与线性拟合图

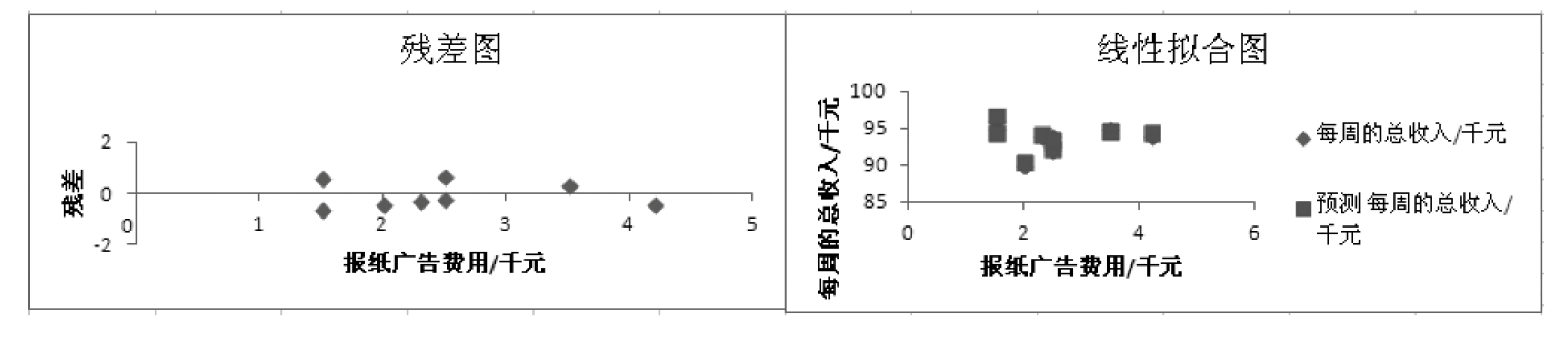


图 3-23　报纸广告费用残差图与线性拟合图

X_1和 X_2系数的检验值 P 值小于 0.05，因此本示例中的电视广告费用和报纸广告费用对收入均有显著性的影响。

3.4 时间序列分析

时间序列是指将某一个变量在不同时间上的各个数据按时间先后顺序排列而形成的序列。时间序列分析的主要目的是根据已有的历史数据对未来进行预测。时间序列分析基于随机过程理论和数理统计学方法,研究数据随时间发展变化的规律性。它包括一般统计分析(如自相关分析、谱分析等),统计模型的建立与推断,以及关于时间序列的最优预测、控制与滤波等内容。经典的统计分析通常假定数据序列具有独立性,而时间序列分析则侧重研究数据序列与时间序列之间的依赖关系。

时间序列由于受到各种偶然或随机因素的影响,具有动态随机变化的性质。从表面看杂乱无章、毫无规律,实际上却具有一定的统计规律性。因此,要想对所研究的时间序列建立适当的模型,首先必须了解时间序列的基本统计特性,从而确保时间序列模型的可靠性,并满足一定的精度。一般可以从时间序列的平稳性、纯随机性和季节性三个方面考虑。

3.4.1 平稳性检验

平稳性是某些时间序列具有的一种统计特征。只有对于平稳的序列才可以运用已知的时间序列模型对其进行分析预测,因此对数据进行平稳性检验是时间序列分析法的关键步骤。

对序列的平稳性有两种检验方法:一种是根据时间序列图显示的特征做出判断的图检验方法;一种是构造检验统计量进行假设检验的方法。通常我们都选用图检验方法来检验序列平稳性,即以时间轴为横轴,变量为纵轴构成时间序列图。

3.4.2 纯随机性检验

如果序列值彼此之间没有任何相关性,则意味着该序列是一个没有记忆的序列,过去的行为对将来的发展没有丝毫影响,这种序列我们称之为纯随机序列。纯随机性检验又称白噪声检验,是专门用来检验序列是否为纯随机序列的一种方法。

判断一个时间序列是否为纯随机序列最简单、最直观的方法是利用自相关函数(ACF)图和偏自相关函数(PACF)图进行分析。自相关函数(ACF)描述时间序列观测值与其过去的观测值之间的线性相关。偏自相关函数(PACF)描述在给定中间观测值的条件下时间序列观测值与其过去的观测值之间的线性相关。在统计分析软件 SPSS 中,ACF 函数图和 PACF 函数图更加直观地给出了显著性水平 $\alpha=0.05$ 时的随机区间,若 ACF 系数和 PACF 系数落入随机区间内则表示该观测值与过去的观测值无关,若几乎所有 ACF 系数与 PACF 系数都落入随机区间,则可认为该序列是纯随机的。

3.4.3 适用性检测

用不同的模型分析数据时,需要一些参数来判断某个模型对这组数据分析的适用性,其中 Akaike 最小信息准则(AIC)和 Schwarz-Bayes 准则(BIC)是常用的两个参数。AIC

和 BIC 共同的特点是在残差最小的情况下，用尽可能少的参数建立模型。在比较两个或多个模型时，一般选用具有最小 AIC 值和 BIC 值的模型。

常用的时间序列分析模型有指数平滑模型、差分运算模型和 ARIMA 模型，下面对这些模型进行简单介绍。

1. 指数平滑模型

指数平滑模型是布朗(Robert G. Brown)提出的。布朗认为时间序列的态势具有稳定性和规则性，所以时间序列可被合理地顺势推延；最近的过去趋势在某种程度上会持续到最近的未来，所以可以把最近的数据设置较大的权重。指数平滑法通过计算指数平滑值，配合一定的时间序列预测模型对事物的未来进行预测。其原理是任一期的指数平滑值都是本期实际观测值与前一期指数平滑值的加权平均。

2. 差分运算模型

差分运算模型是一种非常简便、有效的确定性信息提取方法。Cramer 分解定理在理论上保证了适当阶数的差分一定可以充分提取确定性信息。通常使用差分运算对数据进行平稳化处理。

3. ARIMA 模型

时间序列自回归模型 AR(p)是一种从回归分析中的线性回归发展而来的分析时间序列的方法，它的工作思想是用以前 p 个时间点的值预测未来时间点的值，p 称为自回归项；滑动平均模型 MA(q)是另一种通过历史时间点的值预测未来时间点的值的方法，它的工作思想是用过去 q 个时间点的随机干扰或预测误差的线性组合来表达当前预测值。

如果将自回归模型 AR(p)和滑动平均模型 MA(q)结合，则可得到一个既包含自回归又包含滑动平均的更精确的时间序列分析方法——自回归滑动平均模型 ARMA(p,q)。在实际中，大多数时间序列都是非平稳的(时间序列的平稳性的直观含义是指时间序列没有明显的变化趋势以及没有周期性的有规律的变动)，我们不能直接应用自回归模型 AR(p)，滑动平均模型 MA(q)以及自回归滑动平均模型 ARMA(p,q)通常需要采用差分的方法来处理非平稳的时间序列，这样在自回归滑动平均模型 ARMA(p,q)基础上增加差分处理得到的模型就是自回归积分滑动平均模型 ARIMA(p,d,q)，d 为时间序列成为平稳时所做的差分次数。

自回归积分滑动平均模型 ARIMA 是由博克思(Box)和詹金斯(Jenkins)于 20 世纪 70 年代初提出的著名时间序列预测方法，所以又称为 Box-Jenkins 模型。其原理是将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列，然后使用这个模型根据时间序列的历史值去预测未来值。

3.5　聚类分析

聚类分析属于探索性的数据分析方法。通常，利用聚类分析可以将看似无序的对象进行分组、归类，以达到更好地理解研究对象的目的。聚类结果要求组内对象相似性较高，组间对象相似性较低。在用户研究中，很多问题都可以借助聚类分析来解决，如网站的信息分类问题、网页的点击行为关联性问题以及用户分类问题等。其中，用户分类是最

常见的情况。

聚类分析是根据数据的数值特征对数据进行分类的一种分析方法。与一般的分类算法不同，聚类分析并不能确定数据应该分为几类。聚类分析的目的是将众多的个体先聚集成比较好处理的几个类别或子集，然后再利用判别分析进一步研究各个类别之间的差异。

对一组数据，既可以对变量(指标)进行聚类分析，也可以对观测值进行聚类分析。分析的时候，不一定要事先假定有多少类，也可以完全根据数据自身的规律来分类。一般将变量的聚类分析称为 R 型聚类，而将观测值聚类称为 Q 型聚类。

聚类分析中，比较重要的是两个距离的概念，按照远近程度来聚类是聚类分析法的要义，那么这个远近究竟指什么呢？这里的距离一方面是指点与点之间的距离，另一方面是指类和类之间的距离。点间距离本身有多个定义方式也即多种运算方法，因此只要选择一种算法即可。由一个点组成的类是最基本的类，如果每一类都由一个点组成，那么点间距离就是类间距离。但如果一个类包含不止一个点，那么就需要确定类间距离。类间距离是基于点间距离定义的，如两类之间最近点之间的距离可以作为两类间距离，也可以选用最远点的距离，还可以选择各类之间的中心距离。

聚类分析有多种方法，不同的系统提供了不同的聚类分析法。SPSS 提供了 K-平均值聚类、两步聚类和系统聚类三种聚类方法，但它们的应用范围和优劣势各有不同。

K-平均值聚类(KCA)又称快速聚类，是进行人群细分时最常使用的方法。该方法是单纯应用统计技术根据若干指定变量(应限制为尺度变量)将众多个案分到固定的类别中去。这种方法用于大量(数千)个案的类别划分时非常有效。但该方法可以选择的内容较少，最重要的是选择聚类的数量、迭代的次数以及聚类的中心位置，所以人为经验和判断无形中会起很大作用。KCA 方法本身不仅要求确定分类的类数，而且需要事先确定点，也就是聚类种子。在实际操作中，SPSS 会自动选取种子，然后根据其他点离这些种子的远近对所有点进行分类。再然后，就是将这几类的中心(均值)作为新的基石再分类，如此迭代。

两步聚类是揭示自然类别的探索性工具。该方法的算法与传统聚类技术相比有一些显著的特点：它可以基于类别变量和连续变量来进行聚类；自动选择聚类结果的最佳类别数；具备有效分析大量数据的能力。

如果只拥有少量的个案(少于数百个)，并且想尝试多种聚类方法，测量不同类别之间的差异，则应该尝试使用系统聚类。系统聚类也叫层次聚类(HCA)。当然该方法不仅可以对样本聚类，还可以对变量聚类。这种方法的分类结果取决于对聚类方法、距离测量方法、标准化变量的设置。这种方法不事先确定类数，有多少点就是多少类，它沿着最近的先聚为一类的思想进行合并，直至最后只有一个大类为止。

3.6　可视化数据分析

随着信息化的发展和信息技术在社会与国民经济各个领域的广泛应用，产生了海量的数据信息，面对海量庞杂的数据人们越来越希望能有直观、易懂的方式查看数据。数据

可视化技术应运而生,将数据转换成容易理解的图形,用图形的方式展现数据之间的关系并应用现有的数据对事物未来的发展做出预测。

可视化技术按照其目的可以分成三大类:探索型、验证型、表示型。探索型是指人们预先对数据没有任何认识,通过可视化技术对数据进行分析得到数据的规律与发展趋势后提出关于数据的假设。验证型是人们事先提出针对数据的假设之后应用可视化技术对数据进行分析,验证假设是否合理。表示型指的是应用有效的手段或者技术表示数据。

本节首先介绍几种常用的可视化分析图表,然后介绍几个用这些图表进行分析的示例。

3.6.1　常用的可视化数据展示方法

1. 条形图

条形图又称条状图、柱状图、柱形图,是最常用的图表类型之一。它通过垂直方向或水平方向展示维度字段的分布情况。水平方向的条形图即为一般意义上的条形图,垂直方向的条形图通常称为柱形图。

条形图可以迅速对数据做出比较,一目了然地揭示高低点。如果数值数据能够被归入不同类别,那么条形图就尤为有效,便于快速看清数据中显示的趋势。

条形图适合跨类别比较数据,如按来源站点划分的网站流量、按区域划分的消费比率。

图 3-24 为一个柱状图示例,图 3-25 为一个条形图示例。

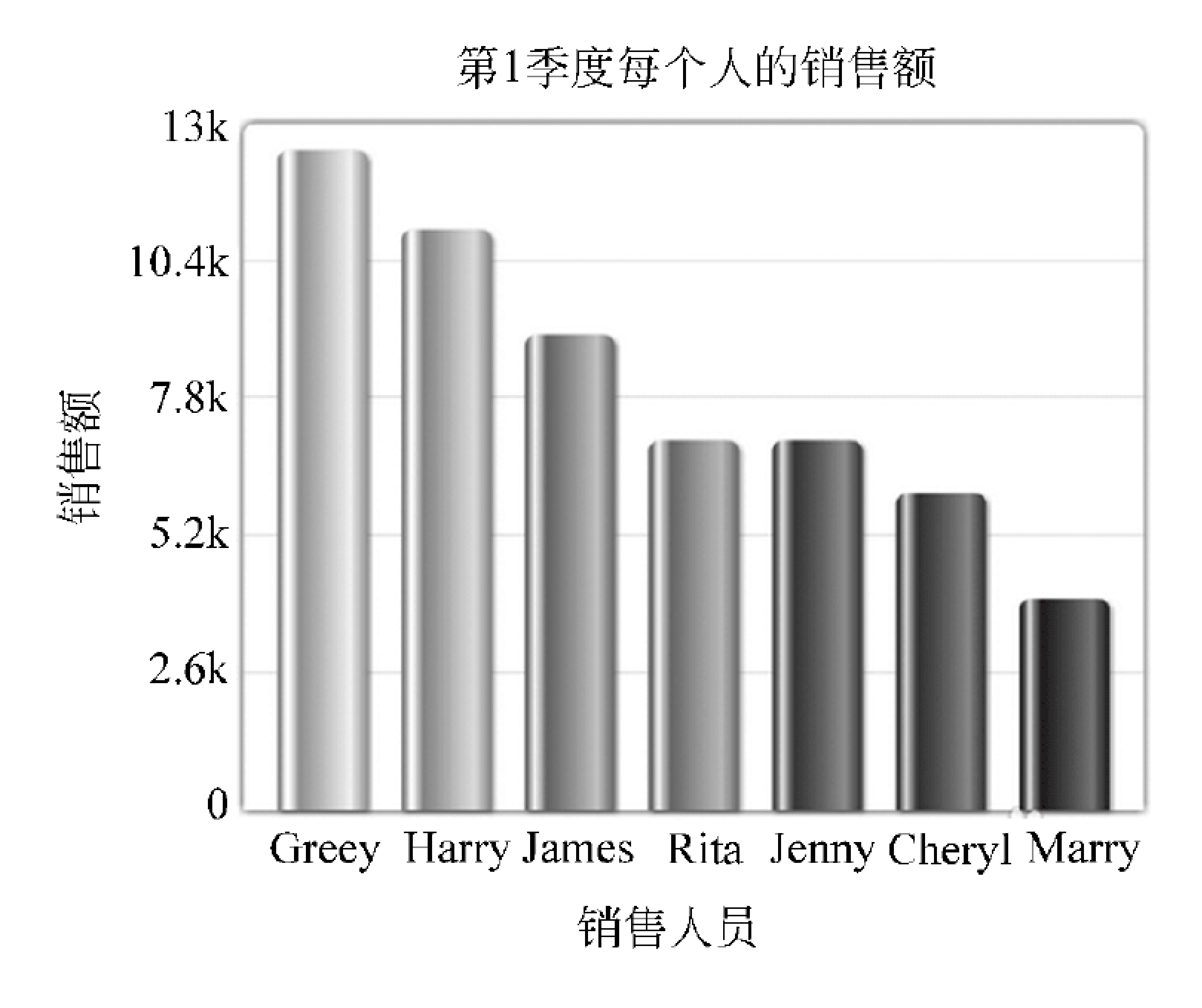


图 3-24　柱状图示例

2. 直方图

直方图(Histogram)又称质量分布图,它与条形图类似,主要区别在于条形图主要用于展示分类数据,直方图主要用于展示数值型数据。

条形图用于展示不同类别的数据时,类别是离散的、较少的,而直方图则是对此类别再进行分组统计。分组的原因可能是类别是连续的,或者类别虽然离散但数量很多,可以

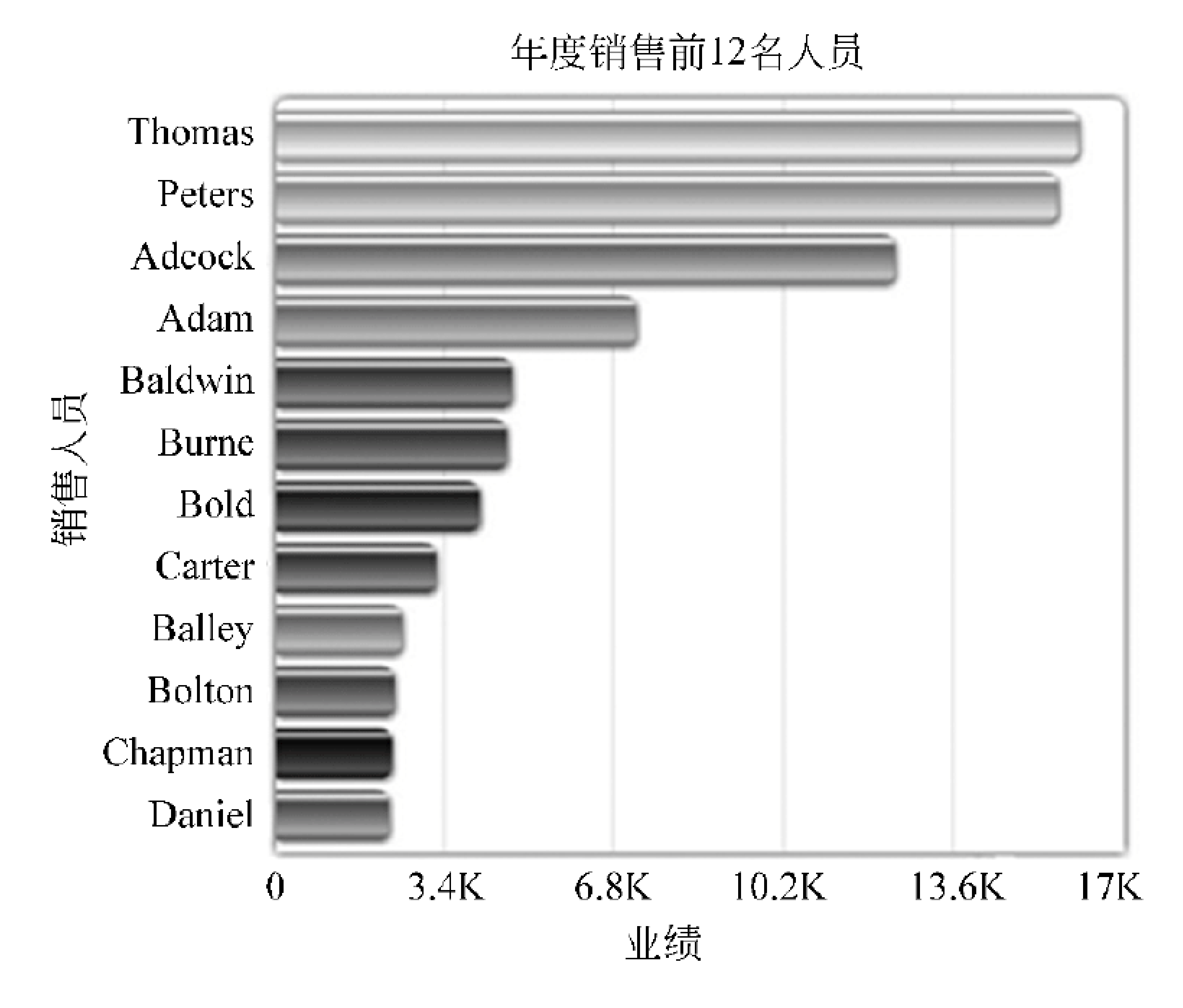


图 3-25　条形图示例

视为近似于连续的。使用直方图分析的样本数据量最好在 50 个以上。

如果希望了解数据的分布情况，如学生考试成绩各分数段情况、产品缺陷频率等，则可使用直方图表示。

如图 3-26 所示为用直方图形式显示学生考试成绩各分数段的人数分布情况。

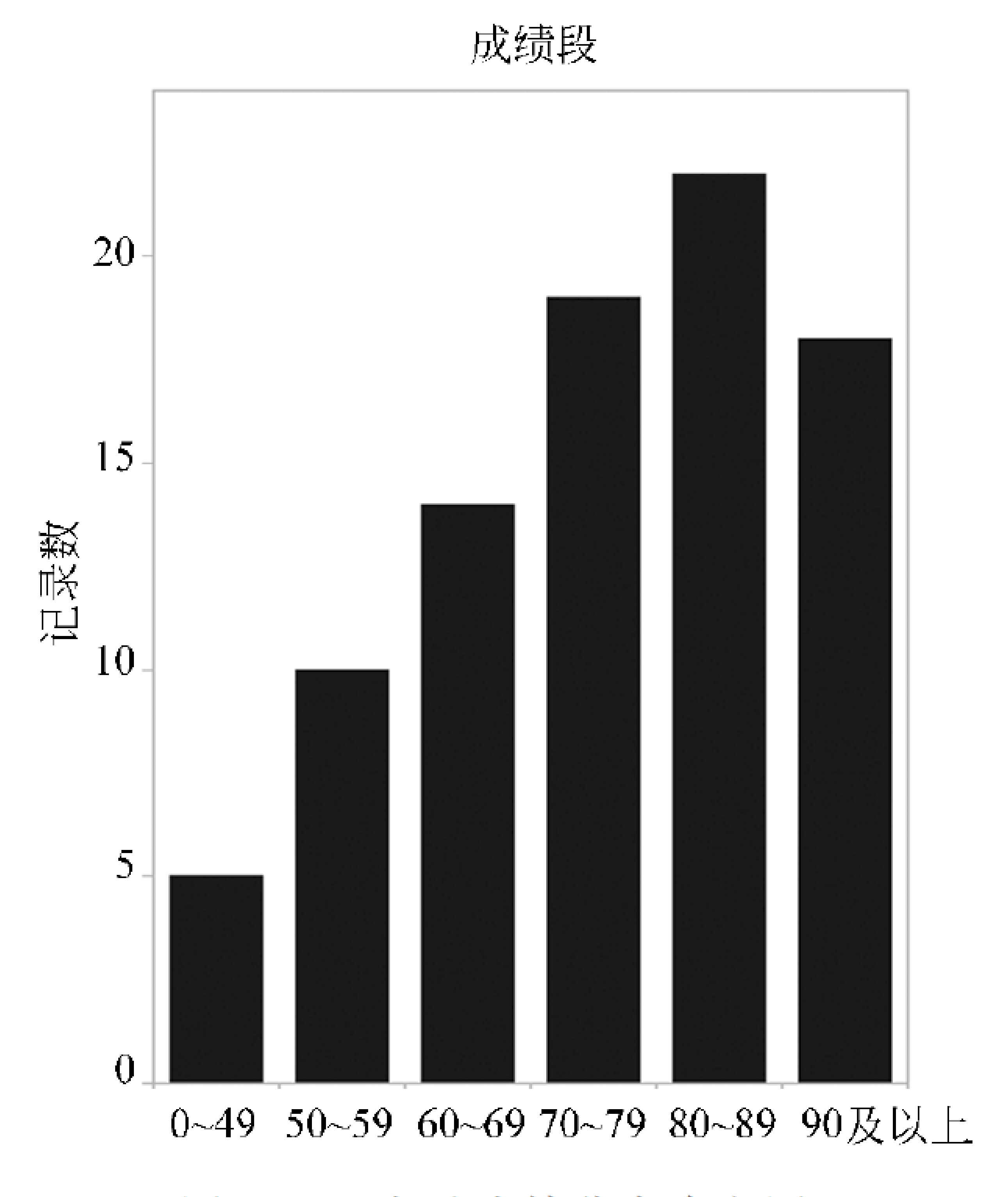


图 3-26　考试成绩分布直方图

3. 折线图

折线图可以显示随时间（根据常用比例设置）而变化的连续数据，因此非常适合显示在相等时间间隔下数据的趋势。在折线图中，类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布。

折线图的主要作用是显示一段时间内的数据的变化趋势，如五年期的股价变化、一个月内的网页浏览量等。

如图 3-27 所示为一个折线图示例。

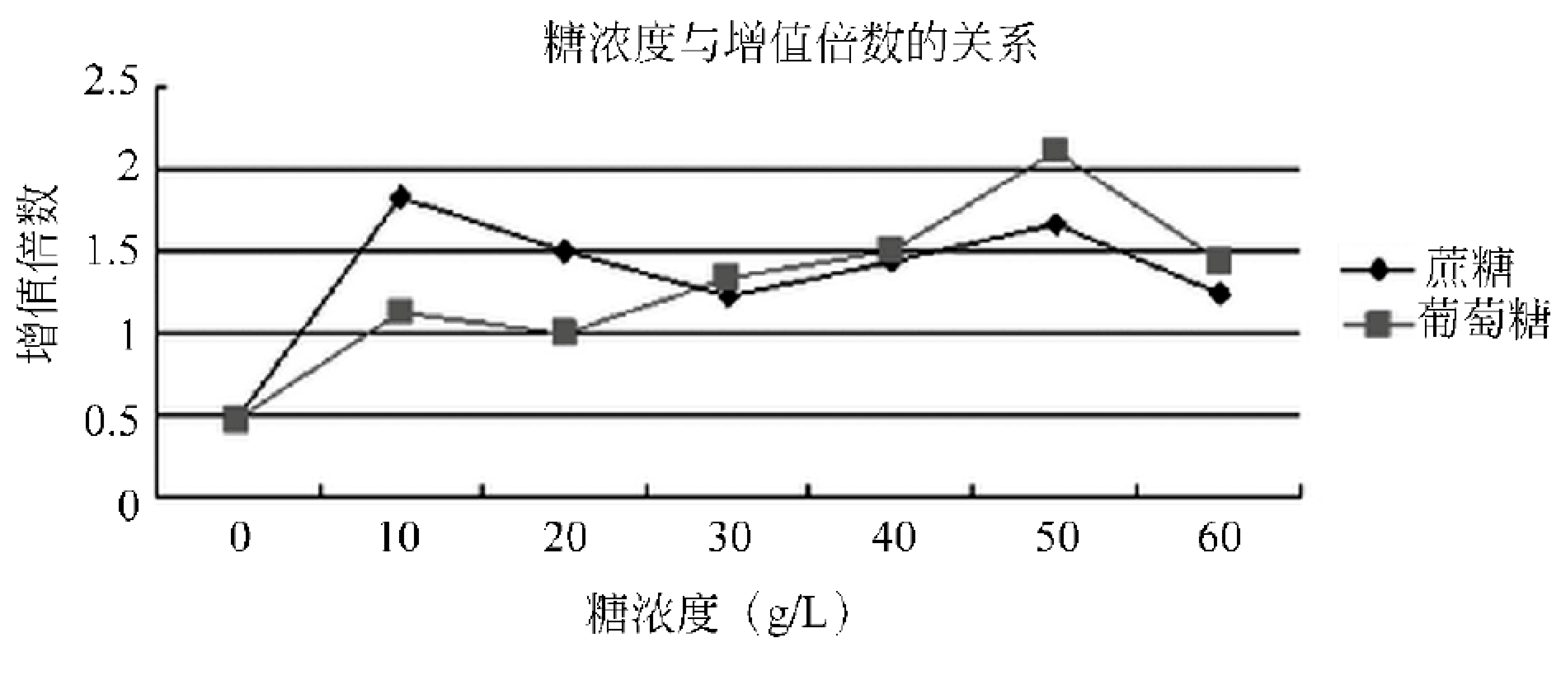


图 3-27　折线图示例

4. 散点图

散点图表示因变量随自变量而变化的大致趋势。一般情况下，散点图用两组数据构成多个坐标点，通过观察坐标点的分布，判断变量间是否存在关联关系以及相关关系的强度。

需要注意的是，相关关系不同于因果关系，相关性表示两个变量同时变化，而因果关系是一个变量导致另一个变量变化。散点图只是一种数据的初步分析工具，能够直观地观察两组数据可能存在什么关系，在分析时如果找到变量间存在可能的关系，则需要进一步确认是否存在因果关系，这需要使用更多的统计分析工具进行分析。

进行相关关系分析时，应使用连续数据，一般在 x 轴（横轴）上放置自变量，在 y 轴（纵轴）上放置因变量，在坐标系上绘制相应的点。散点图的形状可能表现为变量间的线性关系、指数关系或对数关系等。以线性关系为例，散点图一般会包括如下几种典型形状：

（1）正相关：自变量 x 变大时，因变量 y 随之变大；

（2）负相关：自变量 x 变大时，因变量 y 随之变小；

（3）不相关：因变量 y 不随自变量 x 的变化而变化。

如图 3-28 所示为分析收货天数与客户满意度之间的关系的散点图。

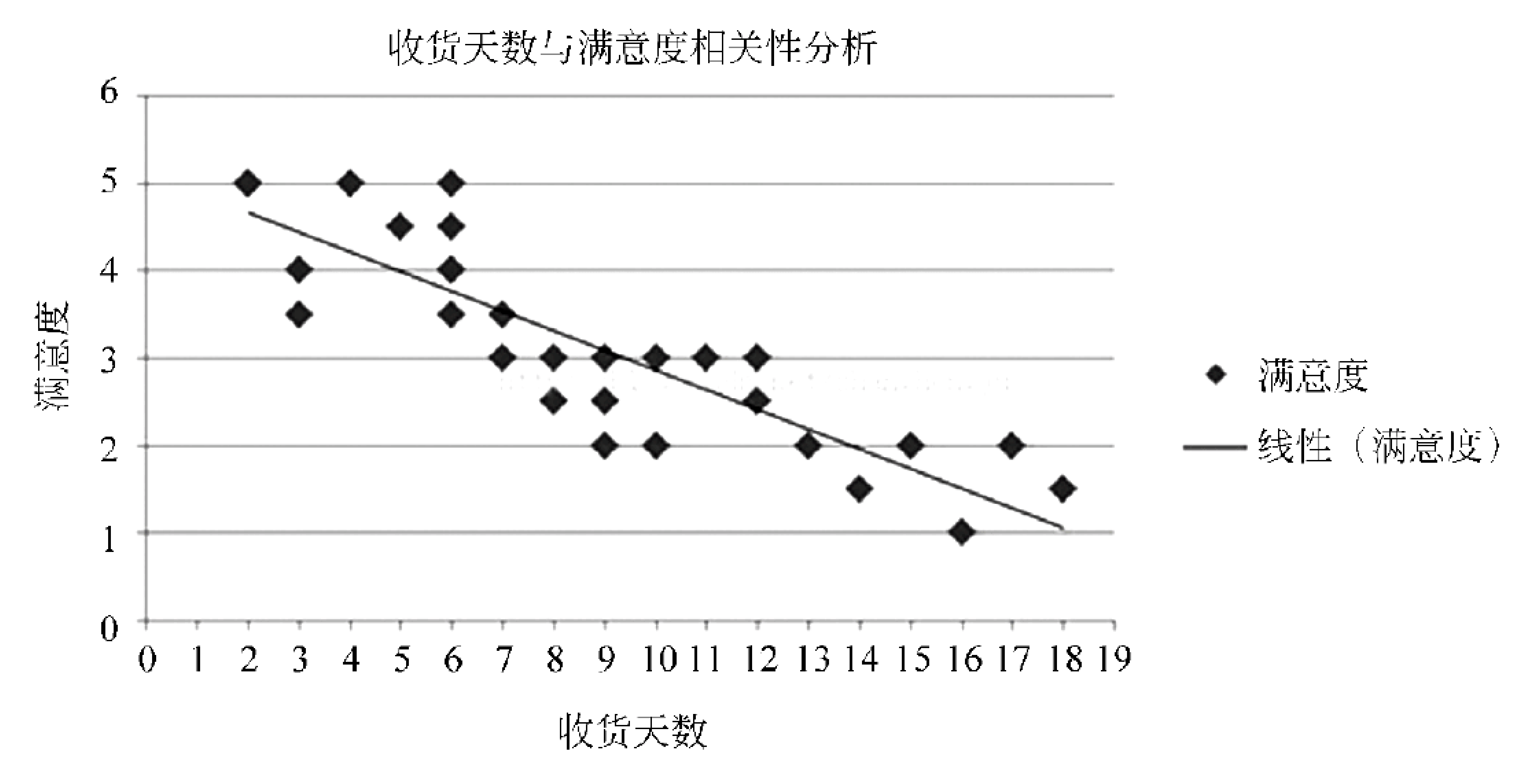


图 3-28　收货天数与客户满意度的相关性分析

5. 气泡图

气泡图不是自成一类的可视化图形,气泡图与散点图类似,不同之处在于,散点图对成组的两个数值进行比较(X 轴和 Y 轴),而气泡图对成组的三个数值进行比较,第三个数值确定气泡数据点的大小。气泡图用圆圈的不同大小揭示数据的意义。

气泡图的特点是具有视觉吸引力,能以非常直观的方式展示数据。

图 3-29 显示了用气泡图形式展示学生考试成绩各分数段的人数情况,气泡越大代表这个分数段的人数越多。

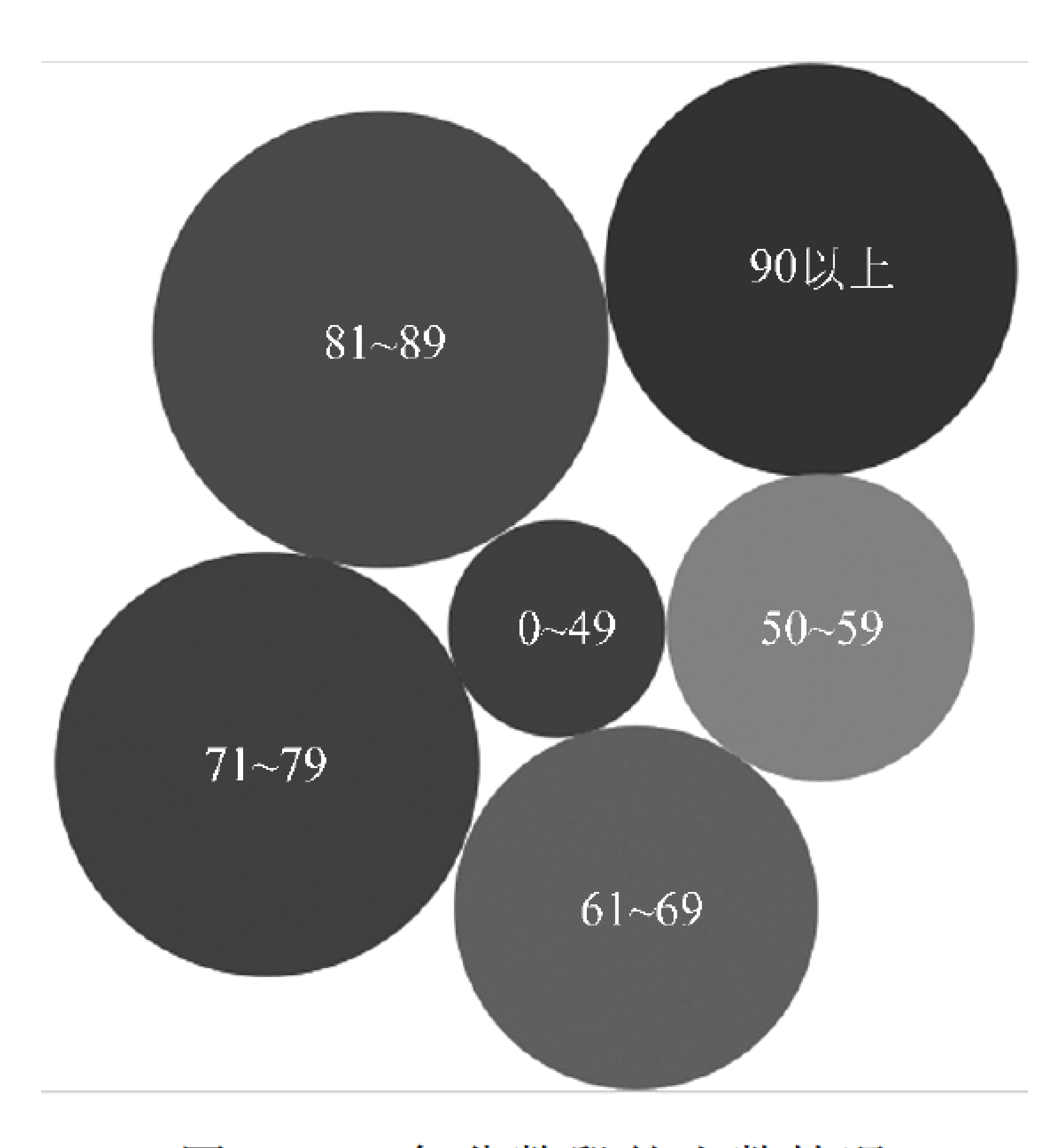


图 3-29　各分数段的人数情况

6. 盒须图

盒须图又称盒式图、箱形图或箱线图,是一种用于显示数据的位置、分散情况、异常值的统计图,因形状如箱子而得名,常应用于品质管理领域。

盒须图上包括 6 个数据节点,将一组数据从大到小排列,分别计算出上限、上四分位数 Q3(也称为第三四分位数)、中位数 Q2、下四分位数 Q1(也称为第一四分位数)、下限,还有一个异常值。

- 中位数:数据按照从大到小的顺序排列,位于中间位置的数,即总观测数的 50% 的数据。
- 第一四分位数:数据按照从大到小的顺序排列,处于总观测数 25% 位置的数据。
- 第三四分位数:数据按照从大到小的顺序排列,处于总观测数 75% 位置的数据。
- 上限:一般情况下,上限 = Q3 + 1.5 × (Q3 − Q1)。也可以人工设置上限值。
- 下限:一般情况下,下限 = Q1 − 1.5 × (Q3 − Q1)。也可以人工设置下限值。
- 异常值:上限和下限之外的数据。

一般来说,上限与第三四分位数之间以及下限与第一四分位数之间的形状称为须状,第三分位数与第一分位数之间的形状称为盒状。盒须图的示意图如图 3-30 所示。

计算四分位数首先要确定 Q1、Q2、Q3 的位置(n 表示数字的总个数):

- Q1 的位置 = $(n+1)/4$
- Q2 的位置 = $(n+1)/2$

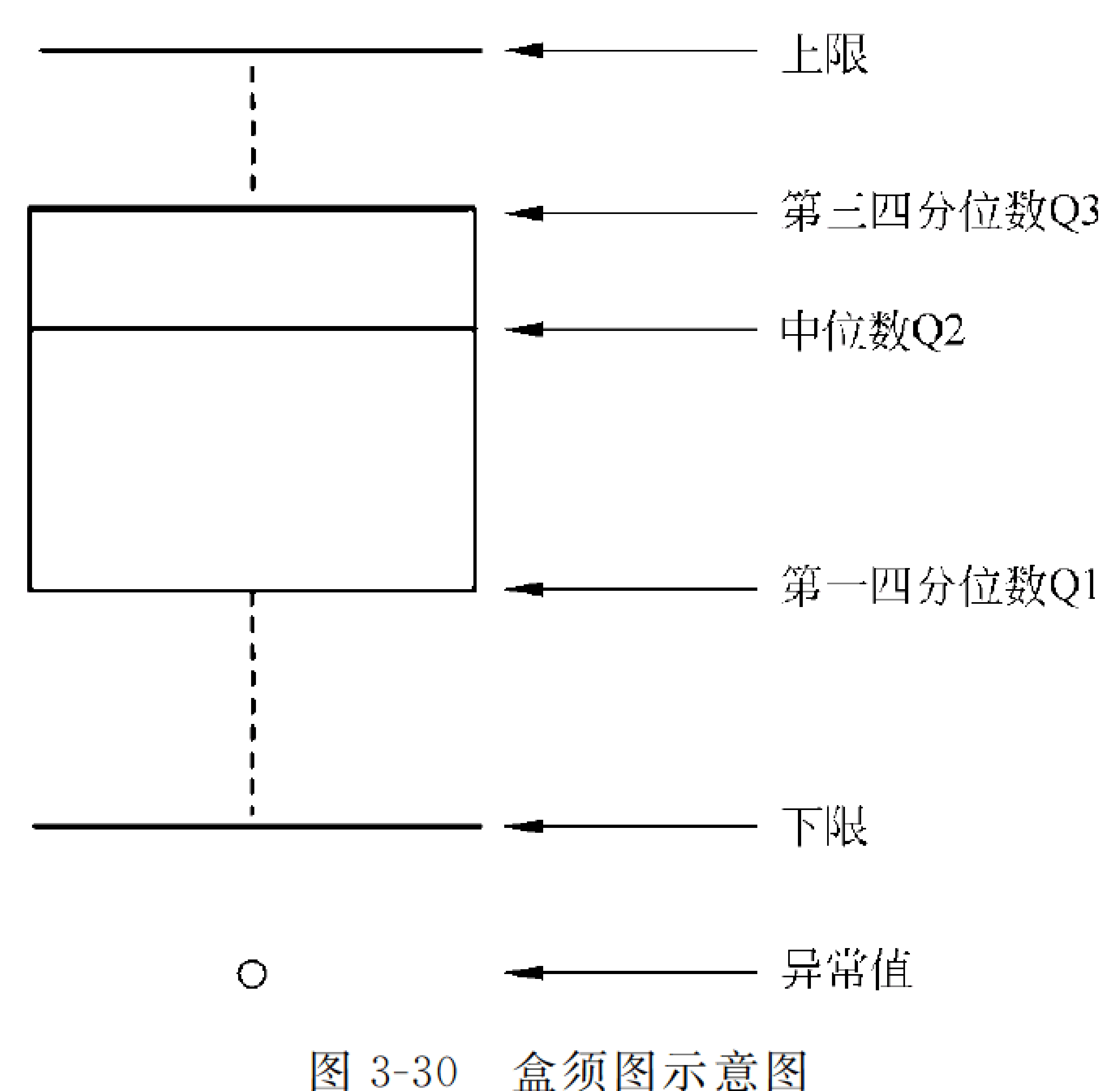


图 3-30 盒须图示意图

- Q3 的位置＝3×(n＋1)/4

盒须图中盒部分的顶部线条是第三四分位数的位置，即 Q3，表示有 75％的数据小于或等于此值；底部线条是第一四分位数的位置，即 Q1，表示有 25％的数据小于此值。整个盒所代表的是数据集中 50％(即 75％～25％)的数据，盒的高度就是这些数据涉及的范围，能够表现出数据的集中程度。

盒须图的美中不足之处在于它不能提供关于数据分布偏态和尾重程度的精确度量，对于批量较大的数据批，盒须图反映的形状信息更加模糊；用中位数代表总体平均水平有一定的局限性等。因此，应用盒须图最好结合均值、标准差、偏度和分布函数等其他统计工具来描述数据批的分布形状。

7. 地图

统计数据是社会经济现象的反映，必然要考虑反映对象的空间位置、空间活动状态及统计对象在空间上的相互作用与影响。地图是一种很好地展示数据空间分布特征的可视化形式。

当数据中包含地理编码、邮政编码、国家名、省/市等位置信息数据时，就可以使用地图来展示数据，如按国家划分的出口目的地、自定义的销售区域等。

可以将地图与其他图形组合起来使用，以便更好地展示信息。例如，将地图与气泡图一起使用，用气泡图展示数据的集中度和不同大小，用地图阐释不同数据点的地理影响。

3.6.2 可视化分析示例

本节介绍柏拉图和散点图的应用，探索可视化技术在实际中的使用。

3.6.2.1 示例 1——柏拉图

柏拉图分析法是 19 世纪的经济学家维尔法度·柏拉图首创的，目的是把一大堆数据

重组，排列成有意义的图表，从而指出问题的原因所在和优次关系。

柏拉图分析的原则是二八原则，即80%的问题是由20%的原因造成的。柏拉图是按照发生频率的大小顺序绘制直方图，将出现的质量问题和质量改进项目按照重要程度依次排列而采用的一种图表。柏拉图常用于分析质量问题、确定影响质量的主要因素、指导纠正措施的实施，以便快速地提升产品质量。

3.6.1节介绍的各种图可以单独使用，也可以组合起来使用。柏拉图就是将柱状图与折线图综合使用的一种数据分析图，是柱状图加折线图的另一种叫法。

柏拉图是分析和寻找影响质量最主要因素的一种工具，其形式是一条分类轴、两条数值轴的坐标图。

- 左边的纵坐标表示频数（如件数、价值等），右边的纵坐标表示频率（百分比）。
- 折线表示累计百分比，这是作图和分析的重点。
- 横坐标表示影响质量的各种因素，按影响程度的大小（即出现频数的多少）从左向右排列。

柏拉图分析法的主要步骤如下。

（1）柏拉图需要使用三列数据：第一列是影响因素名称；第二列是频数，即出现的次数；第三列是各影响因素占的百分比。

（2）对数据按频数进行降序排序。

（3）新生成一列数据：累计百分比，并计算各影响因素到当前行的累计百分比。

（4）使用影响因素、百分比和累计百分比生成簇状柱形图。

（5）需要使用主要横纵坐标轴和次要横纵坐标轴。

（6）更改累计百分比值为折线图。

（7）隐藏次要坐标轴。

1. 示例介绍

表3-4是一家大型铸模公司的数据，该公司制作计算机键盘、洗衣机、汽车和电视机的塑料器件。表中数据是三个月中有缺陷计算机键盘的数据。下面分析当决策制定者着手改进时可以从哪些方面进行改进。

表3-4　三个月中生产的键盘缺陷原因汇总表

原　　因	频　　数	百分比/%
黑点	413	6.53
破损	1 039	16.43
喷射	258	4.08
顶白	834	13.19
划痕	442	6.99
缺料	275	4.35
银条	413	6.53
缩水	371	5.87

续表

原　　因	频　　数	百分比/%
喷雾痕	292	4.62
扭曲变形	1 987	31.42
汇总	6 324	100.01

2. 示例分析

在 Excel 2013 中，对该数据进行如下分析。

(1) 根据表 3-4 中的数据构建 Excel 文件，文件中的数据样式如图 3-31 所示。图中数据是按百分比从大到小排列的。

	A	B	C	D
1	原因	频数	百分比%	累计百分比%
2	扭曲变形	1987	31.42	
3	破损	1039	16.43	
4	顶白	834	13.19	
5	划痕	442	6.99	
6	黑点	413	6.53	
7	银条	413	6.53	
8	缩水	371	5.87	
9	喷雾痕	292	4.62	
10	缺料	275	4.35	
11	喷射	258	4.08	
12	总计	6324	100.01	
13				

图 3-31　键盘缺陷原因汇总表

(2) 计算表中各个累计百分比。选中单元格 D2，输入如下公式并按回车键，如图 3-32 所示。

```
SUM($ C$ 1:$ C2)
```

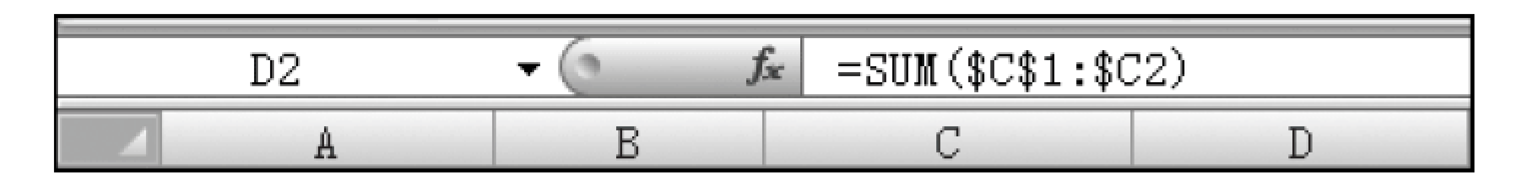


图 3-32　计算累计百分比

选中单元格 D2，移动鼠标至 D2 右下角直到出现一个小的十字架，按住鼠标左键并向 D3 单元格方向移动，一直拖放到 D11 单元格。这样 D3 单元格是 D2 和 C3 两个单元格值的累加，D4 单元格是 D3 和 C3 两个单元格中值的累加，依此类推。最终结果如图 3-33 所示。

(3) 选中单元格 A1 到 A11、C1 到 C11、D1 到 D11，然后在“插入”功能区中单击“插入柱状图”图标，在“二维柱形图”中选择第一个“簇状柱形图”(如图 3-34 所示)，生成如图 3-35 所示的柱形分析图。

(4) 添加次坐标。鼠标右键单击柱形图中“累计百分比”中任意一个矩形条，在弹出的菜单中选择“设置数据系列格式”，在弹出的“设置数据系列”窗格中选择“次坐标轴”，如图 3-36 所示。

	A	B	C	D	E
1	原因	频数	百分比%	累计百分比%	
2	扭曲变形	1987	31.42	31.42	
3	破损	1039	16.43	47.85	
4	顶白	834	13.19	61.04	
5	划痕	442	6.99	68.03	
6	黑点	413	6.53	74.56	
7	银条	413	6.53	81.09	
8	缩水	371	5.87	86.96	
9	喷雾痕	292	4.62	91.58	
10	缺料	275	4.35	95.93	
11	喷射	258	4.08	100.01	
12	总计	6324	100.01		
13					

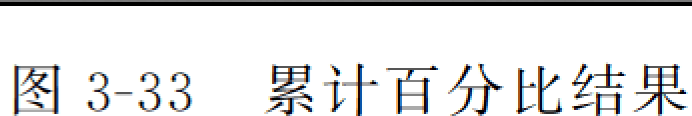


图 3-33　累计百分比结果

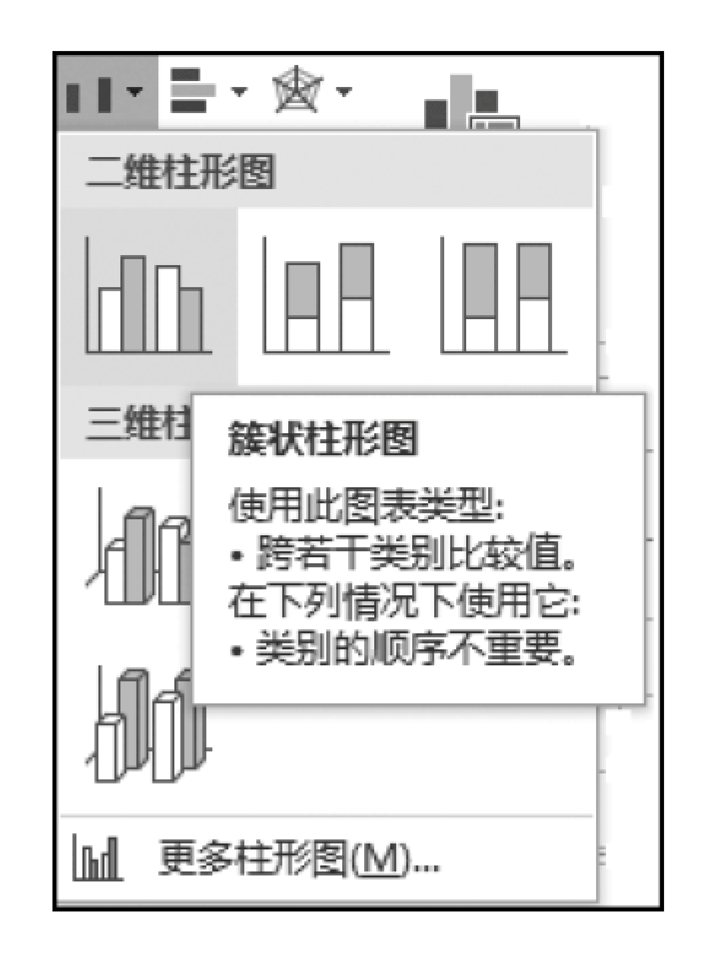


图 3-34　选择“簇状柱形图”

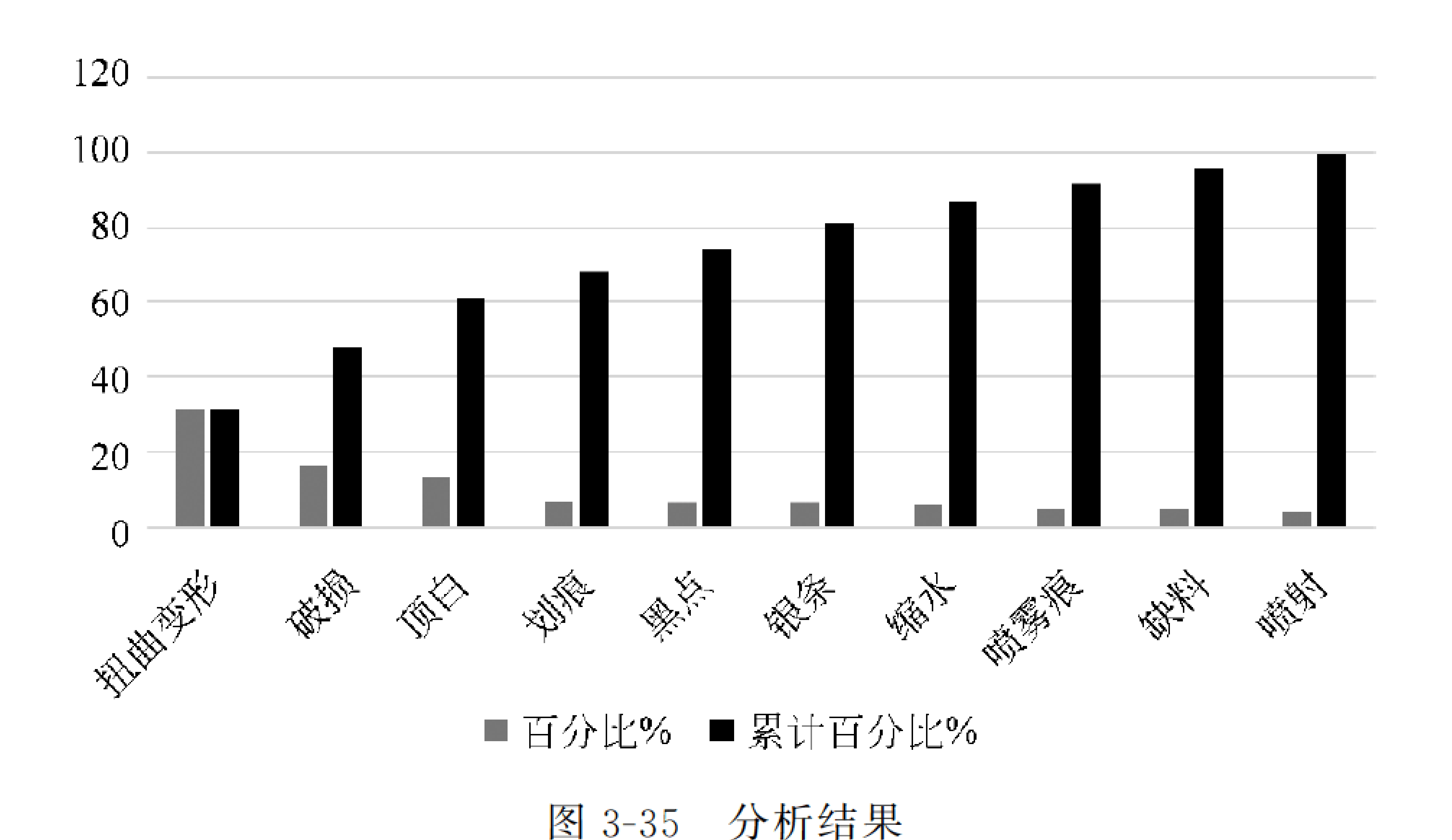


图 3-35　分析结果

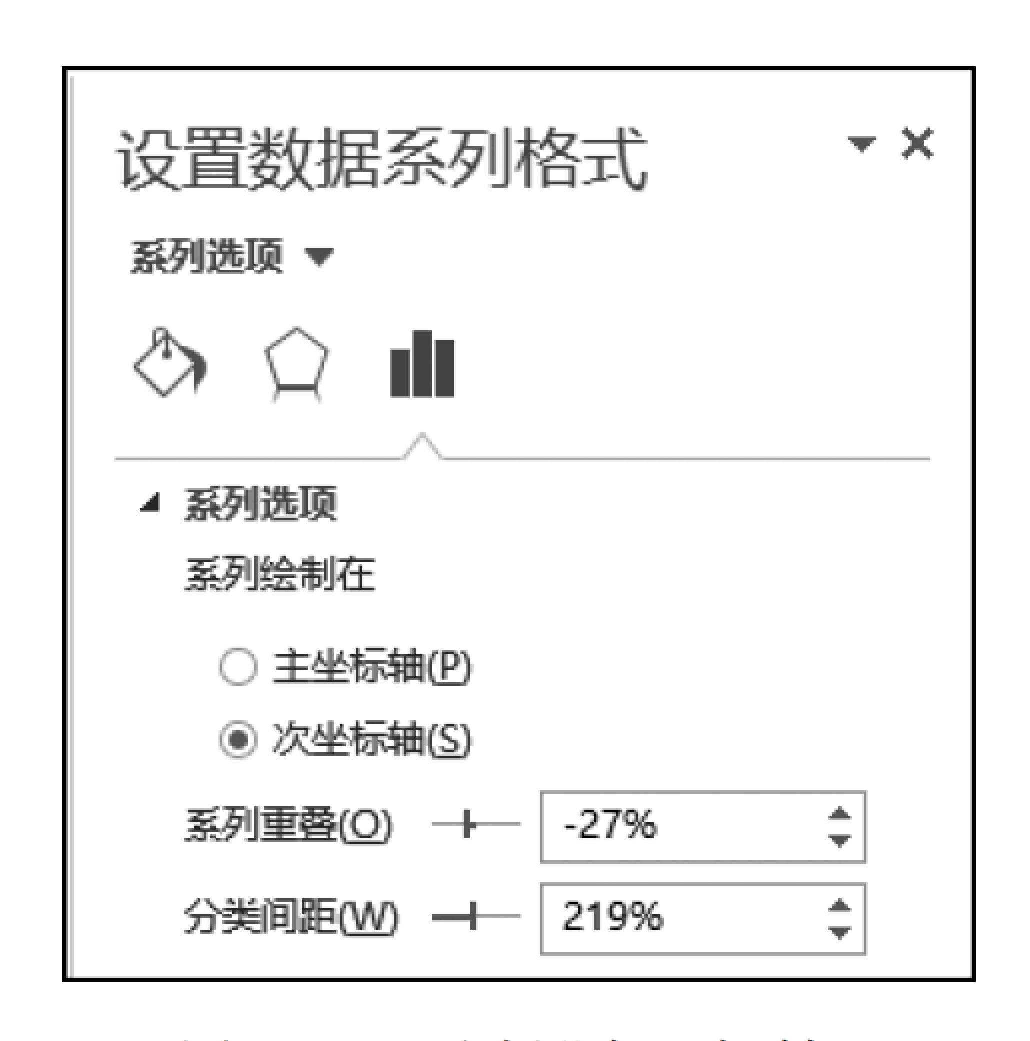


图 3-36　选择“次坐标轴”

（5）关闭“设置数据系统格式”窗格，如图 3-35 所示的柱状图将变为如图 3-37 所示的形式。

（6）将“累计百分比”绘制成曲线。鼠标右键单击柱形图中“累计百分比”中任意一个矩形条，在弹出的菜单中选择“更改系列图标类型”，在弹出的“更改图标类型”窗口中，选择“簇状柱形图-次坐标轴上的折线图”，则图的形式变为如图 3-38 所示的形式。

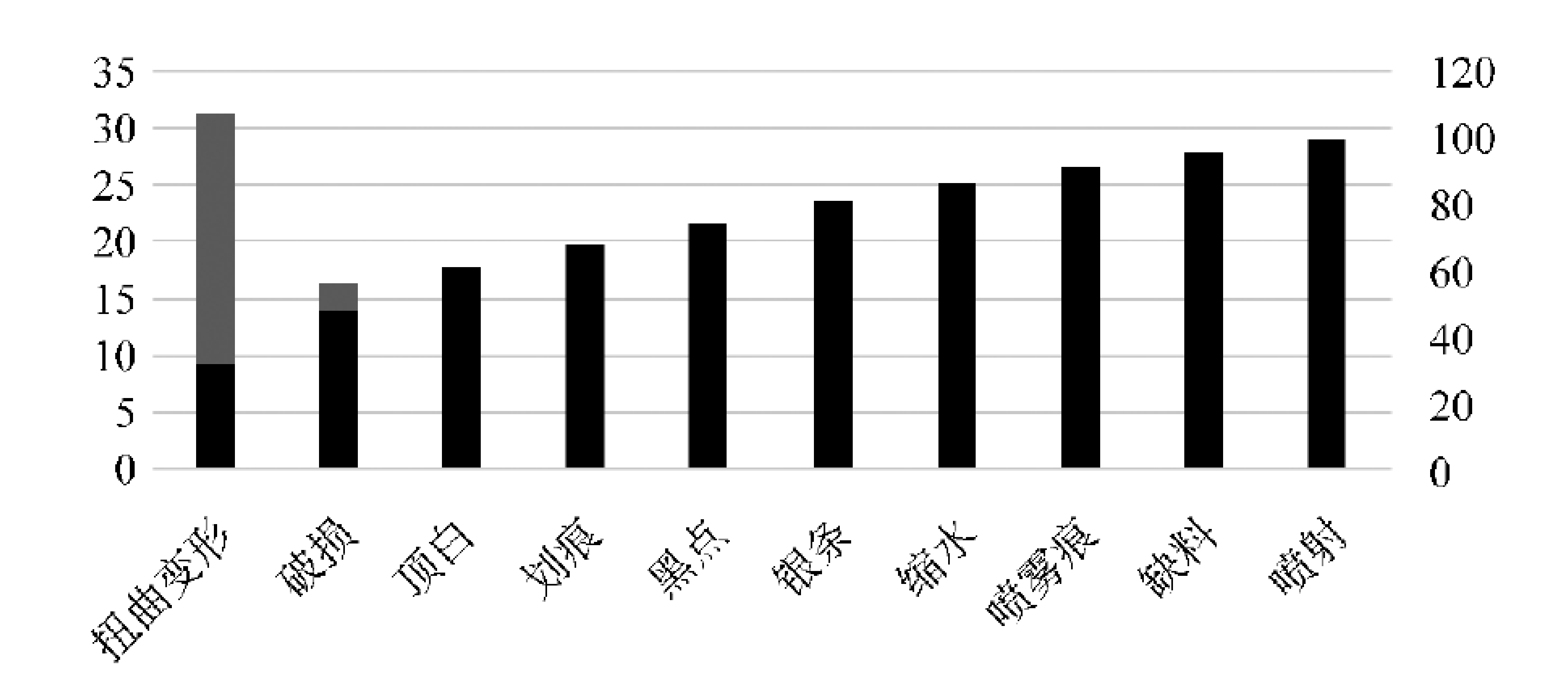


图 3-37　设置好“次坐标轴”后的柱图形式

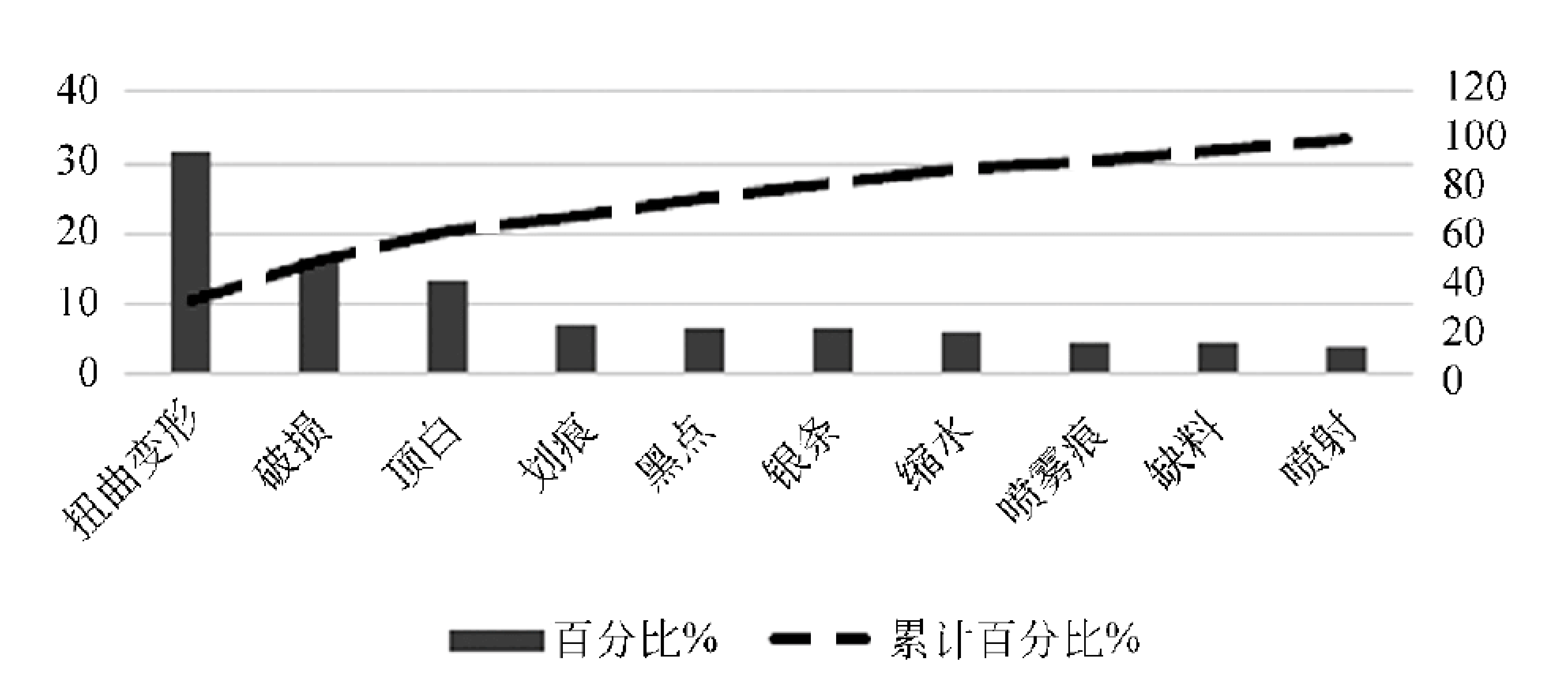


图 3-38　将累计百分比绘制为曲线后的柱图形式

(7) 若要在折线上显示数据，则可在折线上右击鼠标，然后从弹出的菜单中选择“添加数据标签”→“添加数据标签”。

(8) 隐藏次要坐标轴。在折线上右击鼠标，在弹出的菜单中选择“更改数据系列格式”，在出现的“设置数据系列格式”窗格(如图 3-36 所示)中，选中“主坐标轴”。

最终生成的柏拉图如图 3-39 所示。

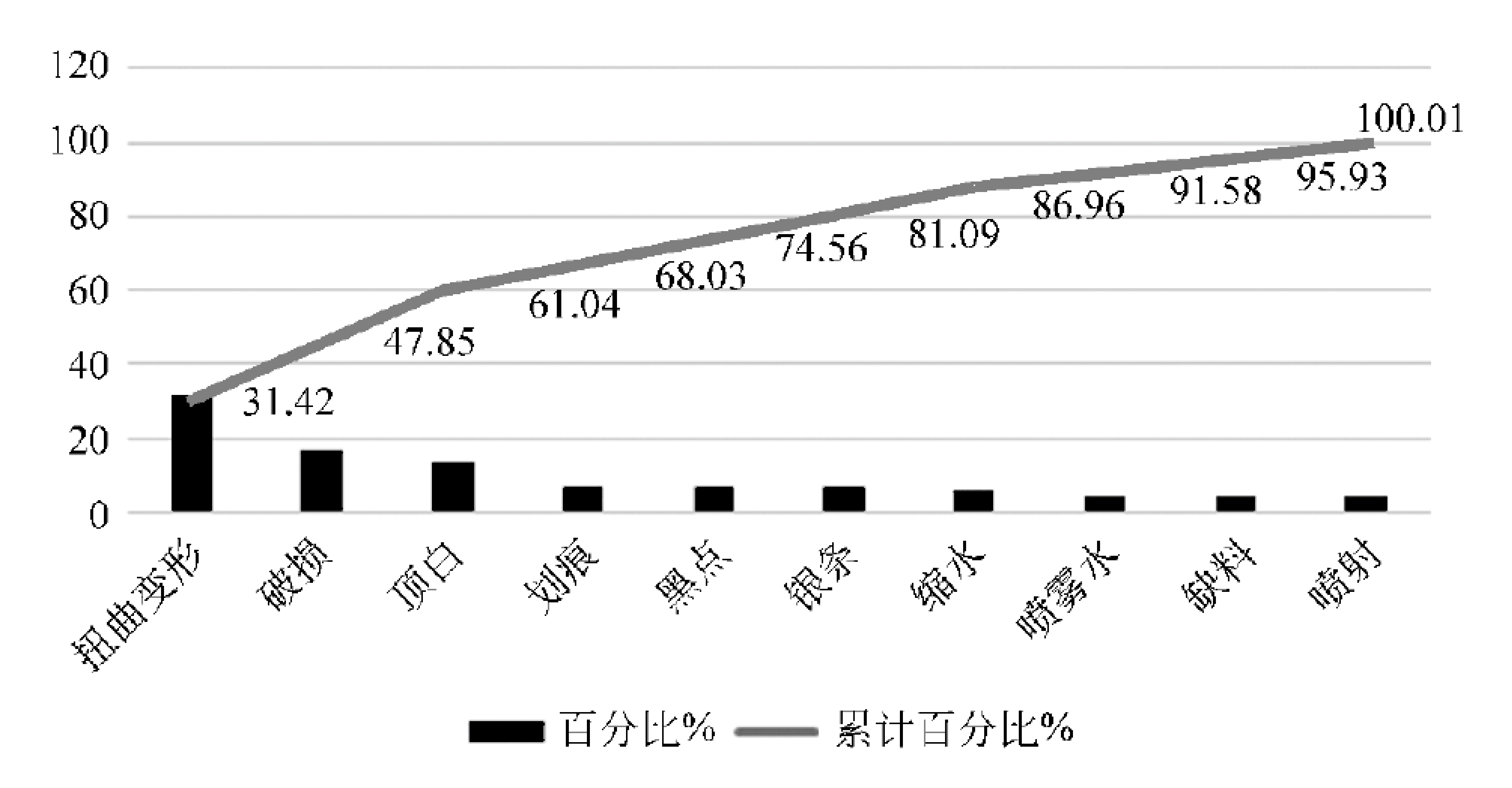


图 3-39　最终生成的柏拉图

3. 结果分析

柏拉图在分析问题的主要因素与次要因素时采用的是二八原则。通过上述生成的柏拉图可知，扭曲变形、破损、顶白、划痕、黑点、银条六个因素占据了约80%，是键盘产生缺陷的主要原因，在六个因素中，扭曲变形、破损、顶白占据了约60%。因此，为了改善键盘的缺陷问题，生产者可以努力减少由扭曲变形、破损、顶白引起的缺陷来获得最大收益，再努力减少划痕、黑点、银条缺陷以进一步提升收益。

3.6.2.2 示例2——散点图

散点图是数据点在直角坐标系平面上的分布图，常常用于判断变量之间是否存在某种关联以及数据的发展趋势。根据散点图中各个变量之间变化的大致趋势，可以选择合适的函数对数据点进行拟合，从而对后续数据的发展进行预测。在选择进行拟合的趋势线时，可以参考参数 R 的平方值，参数越接近1，表示趋势线的拟合程度越高，趋势预测也就越精确。

散点图是用来判断两个变量之间的相互关系的常用工具，一般情况下，散点图用两组数据构成多个坐标点，通过观察坐标点的分布，判断变量间是否存在关联关系，以及相关关系的强度。

进行相关关系分析时，应使用用连续数据，一般在 x 轴(横轴)上放置自变量，y 轴(纵轴)上放置因变量，在坐标系上绘制出相应的点。散点图的形状可能表现为变量间的线性关系、指数关系或对数关系等。以线性关系为例，散点图一般包括如下几种典型形状。

- 正相关：自变量 x 变大时，因变量 y 随之变大；
- 负相关：自变量 x 变大时，因变量 y 随之变小；
- 不相关：因变量 y 不随自变量 x 的变化而变化。

1. 示例介绍

某网站统计了客户收货天数和满意度结果，满意度最高为5分，最低为1分。数据如表3-5所示。下面分析客户收货天数与满意度之间的关联关系。

表3-5 客户收货天数与满意度

收货天数	满意度
6	4.5
12	3
8	3
6	5
18	1.5
7	3.5
3	4

续表

收货天数	满意度
8	2.5
11	3
2	5
12	2.5
15	2
6	4
9	2
2	5
10	2
4	5
13	2
14	1.5
9	3
7	3
3	3.5
6	4
5	4.5
16	1
9	2.5
6	3.5
10	3
17	2

2. 示例分析

在 Excel 2013 中，对该数据进行如下分析。

(1) 数据准备。将如表 3-5 所示的数据录入 Excel 文件中，文件中的数据如图 3-40 所示(图中只展示了部分数据)。

(2) 绘制散点图。选中 A1:B30 区域，在“插入”功能区的“图表”模块中单击“散点图”，选择“仅带数据标记的散点图”图标(如图 3-41 所示)，此时 Excel 中出现初步的散点图，如图 3-42 所示。

	A	B
1	收货天数	满意度
2	6	4.5
3	12	3
4	8	3
5	6	5
6	18	1.5
7	7	3.5
8	3	4
9	8	2.5
10	11	3
11	2	5

图 3-40　收货天数与满意度数据



图 3-41　选择“仅带数据标记的散点图”



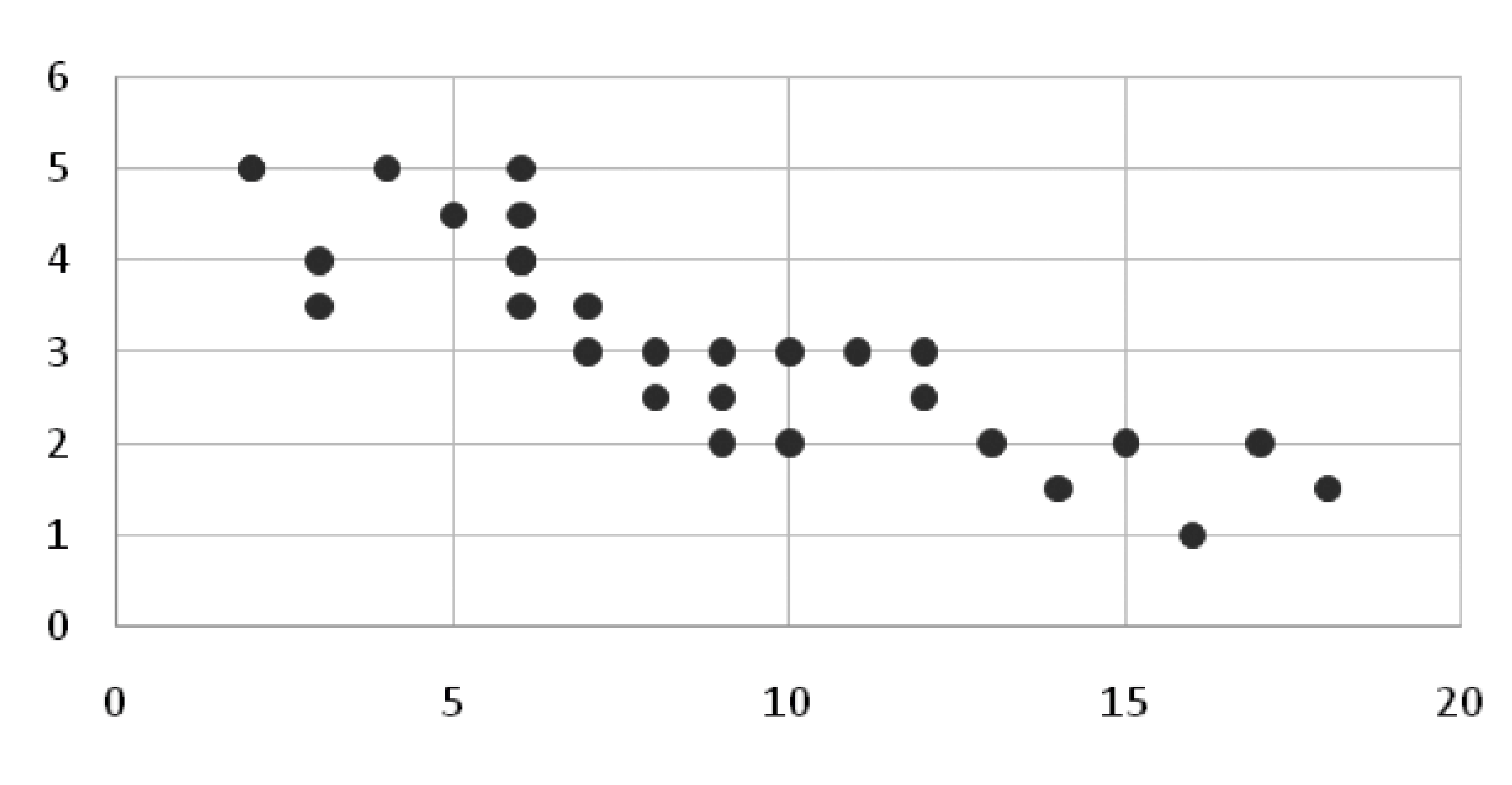


图 3-42　客户满意度散点图

(3) 修改散点图坐标标题。选中散点图，单击右边随之出现的加号图标[+]，在出现的“图表元素”列表框中勾选“坐标轴标题”(如图 3-43 所示)，此时生成带坐标轴标题的散点图，如图 3-44 所示。在图 3-44 中将横、纵坐标的标题修改为“收货天数”和“客户满意度”，如图 3-45 所示。



图 3-43　勾选“坐标轴标题”

(4) 删除散点图的网格线。如果不希望散点图上有网格线，则可将其删除。我们这里删除纵向网格线。选中图 3-45 中的任意一条纵向网格线，单击鼠标右键，在弹出的列表中选择“删除”，结果如图 3-46 所示。

(5) 添加趋势线。选中散点图中的某个数据点，单击鼠标右键，在出现的列表选项中选择“添加趋势线”，出现如图 3-47 所示的“设置趋势线格式”窗格，在此窗格中勾选“显示公式”和“显示 R 平方值”。

关闭“设置趋势线格式”窗格，散点图样式如图 3-48 所示。

(6) 选择最优趋势线。在步骤(5)中，选择不同的趋势线会得到不同的 R^2 值。选择指数时 $R^2=0.758$，线性时 $R^2=0.748\ 4$，对数时 $R^2=0.721\ 2$，多项式时 $R^2=0.764\ 9$，幂

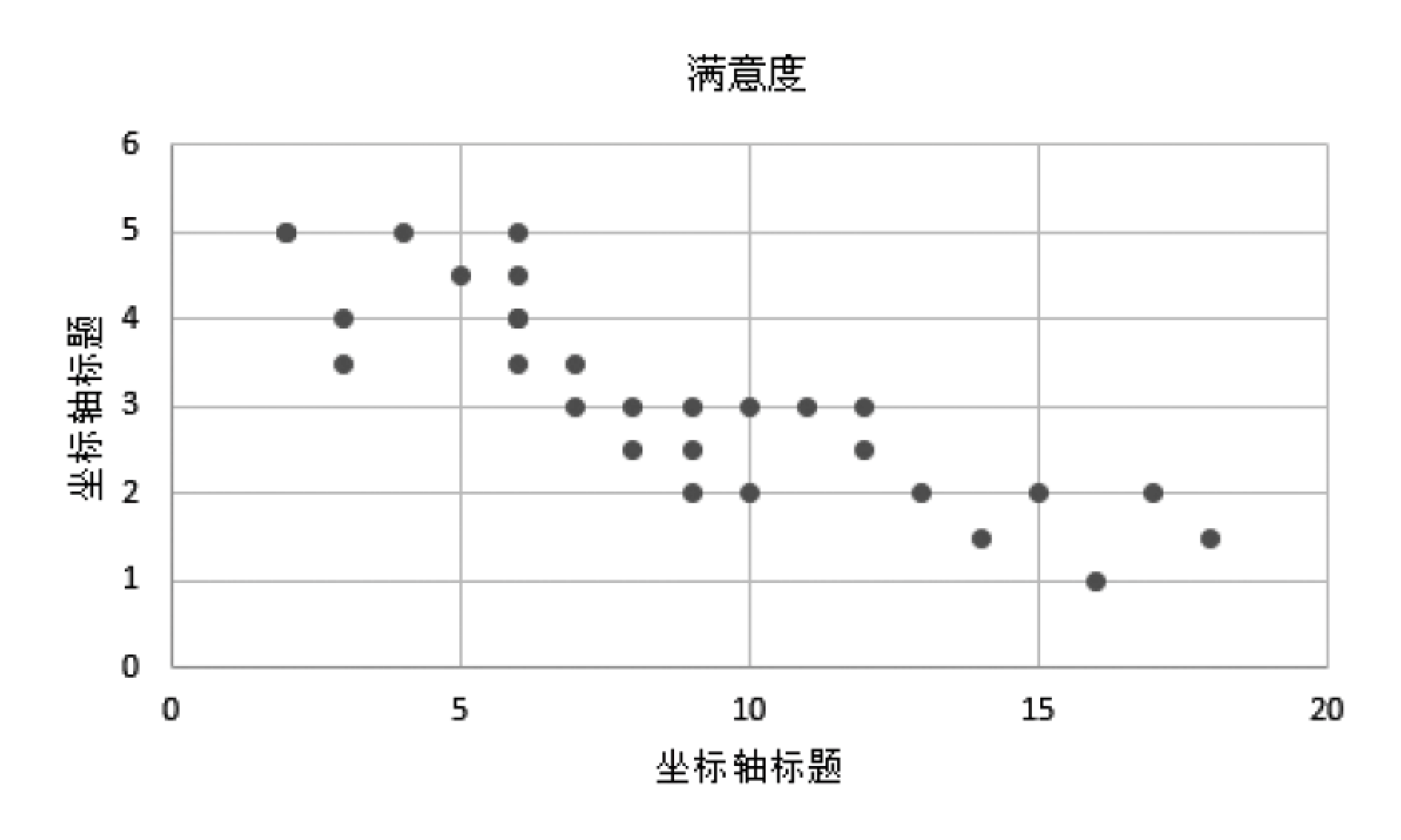


图 3-44　带坐标轴标题的散点图

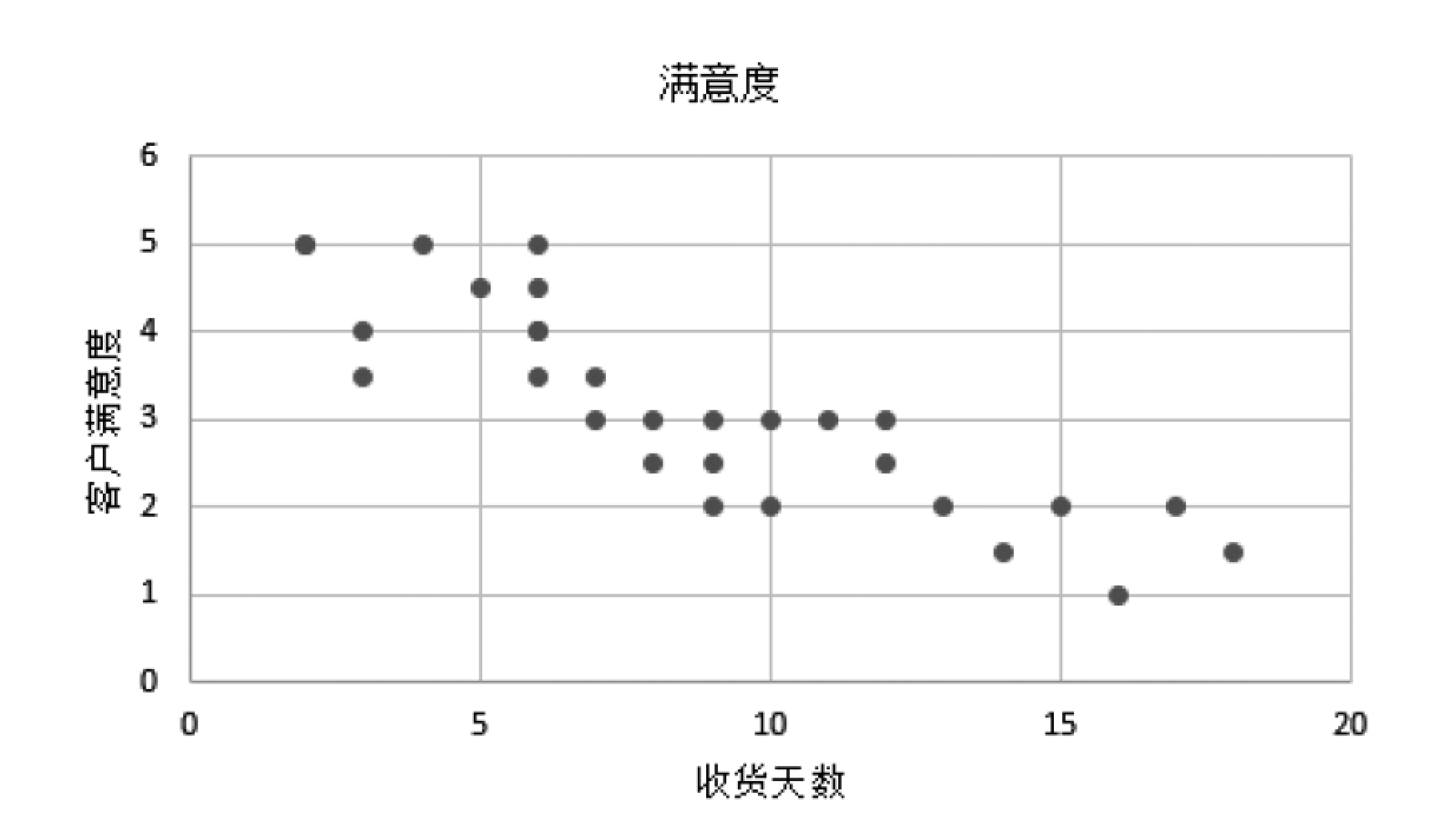


图 3-45　设置好坐标轴标题的散点图

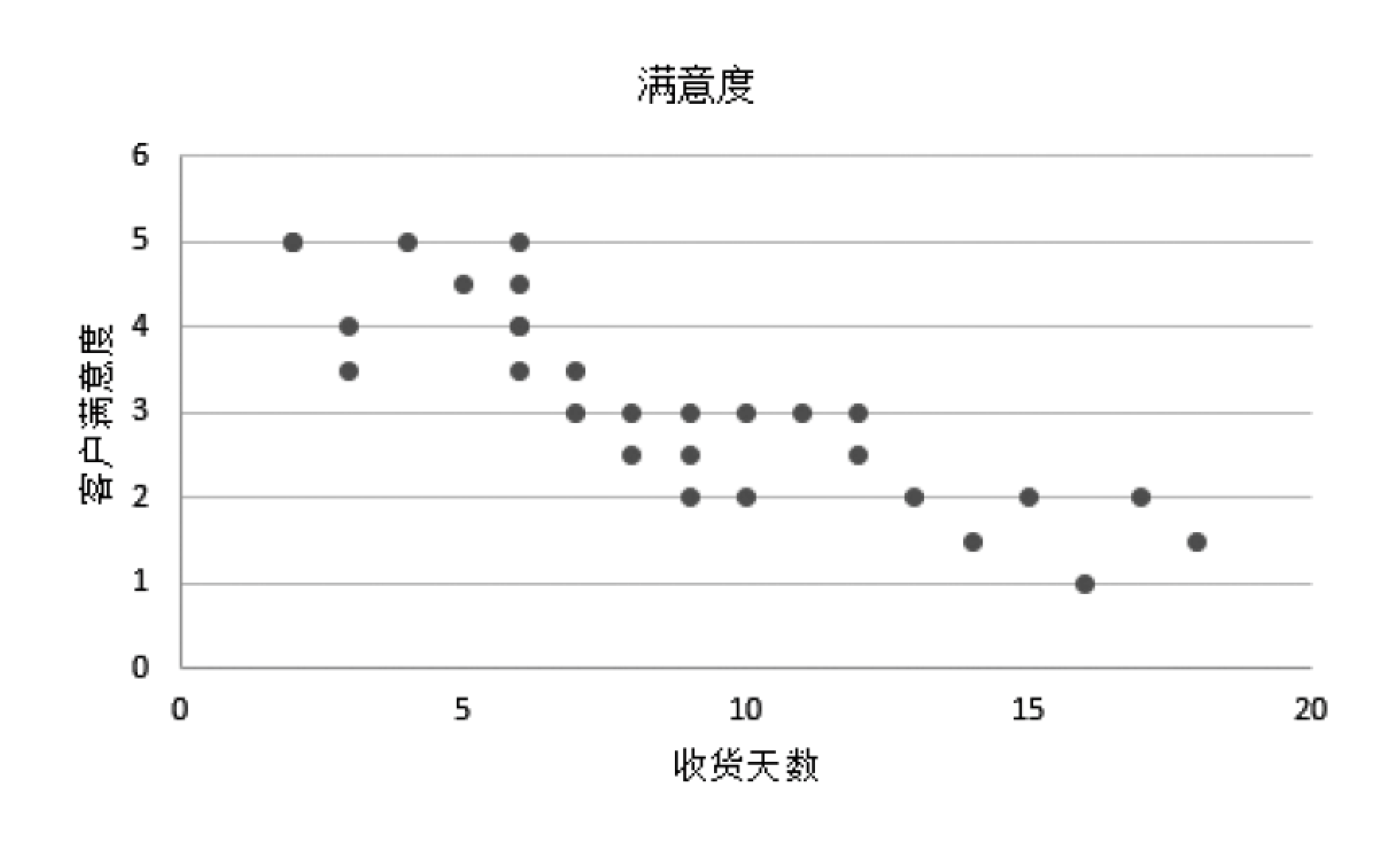


图 3-46　删除了纵向网格线的散点图

时 $R^2=0.6675$。可以看出选择多项式趋势线时，R^2 的值最大，因此在这里选择使用多项式趋势线。最终的散点图如图 3-49 所示。



图 3-47 “设置趋势线格式”窗格

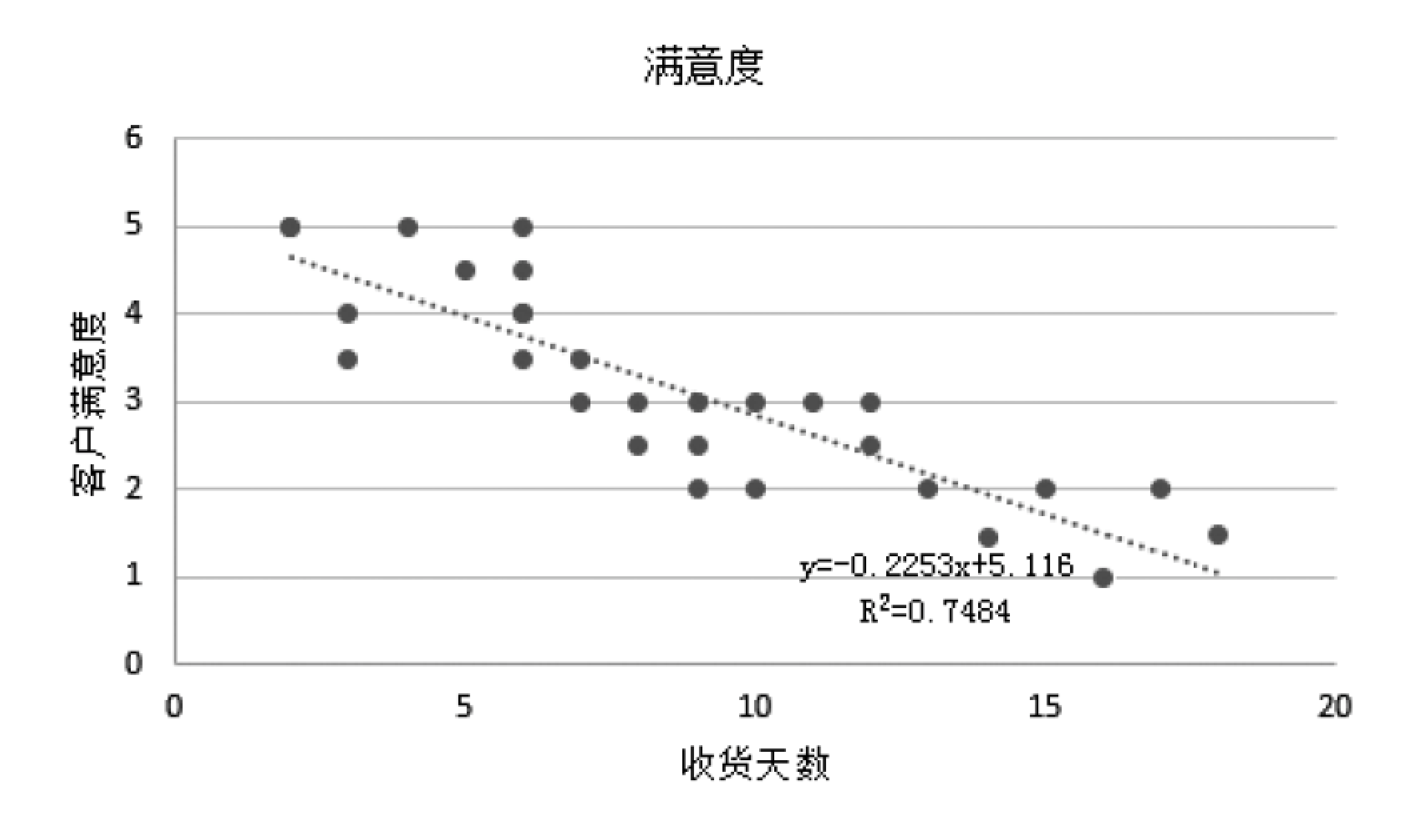


图 3-48 添加了趋势线的散点图

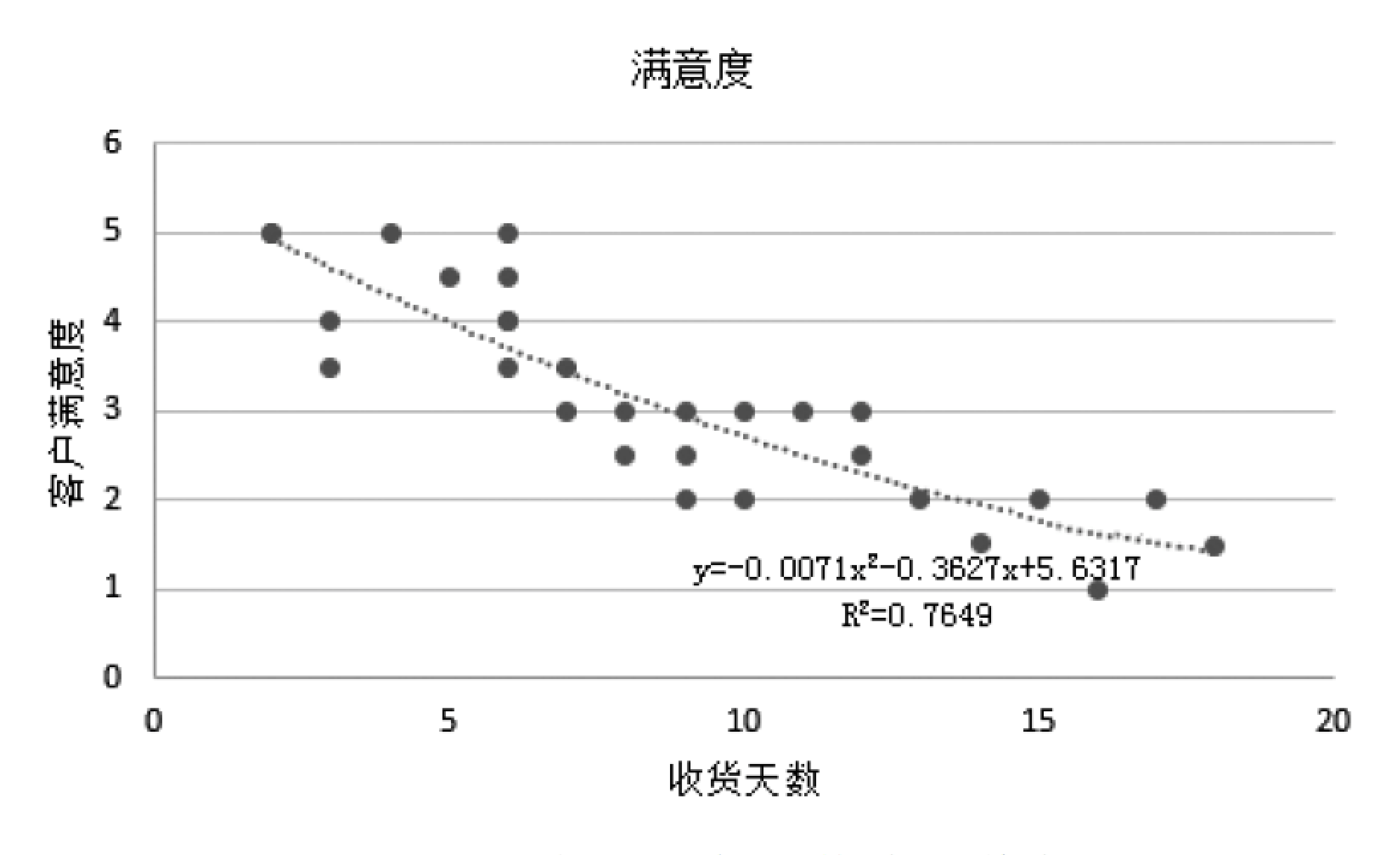


图 3-49 选择了最优趋势线的散点图

说明：在利用散点图做相关分析时，可以添加趋势线。添加趋势线后 Excel 会同时显示回归方程和 R 平方。其中 R 是相关系数，R 平方是决定系数。一般来说，R 的值越高越好。

3. 结果分析

通过分析散点图可以发现，收货天数和客户满意度存在负相关关系，收货天数越长，客户满意度越低。

3.7 环境准备

打开 Excel 2013，单击"数据"菜单，如果在工具栏的最右边没有出现"数据分析"图标 数据分析，则可通过如下步骤将其加到工具栏中。

1. 打开 Excel 2013 文件，在"文件"菜单上单击鼠标左键，在弹出的菜单中选择"选项"命令，弹出"Excel 选项"窗口，如图 3-50 所示。

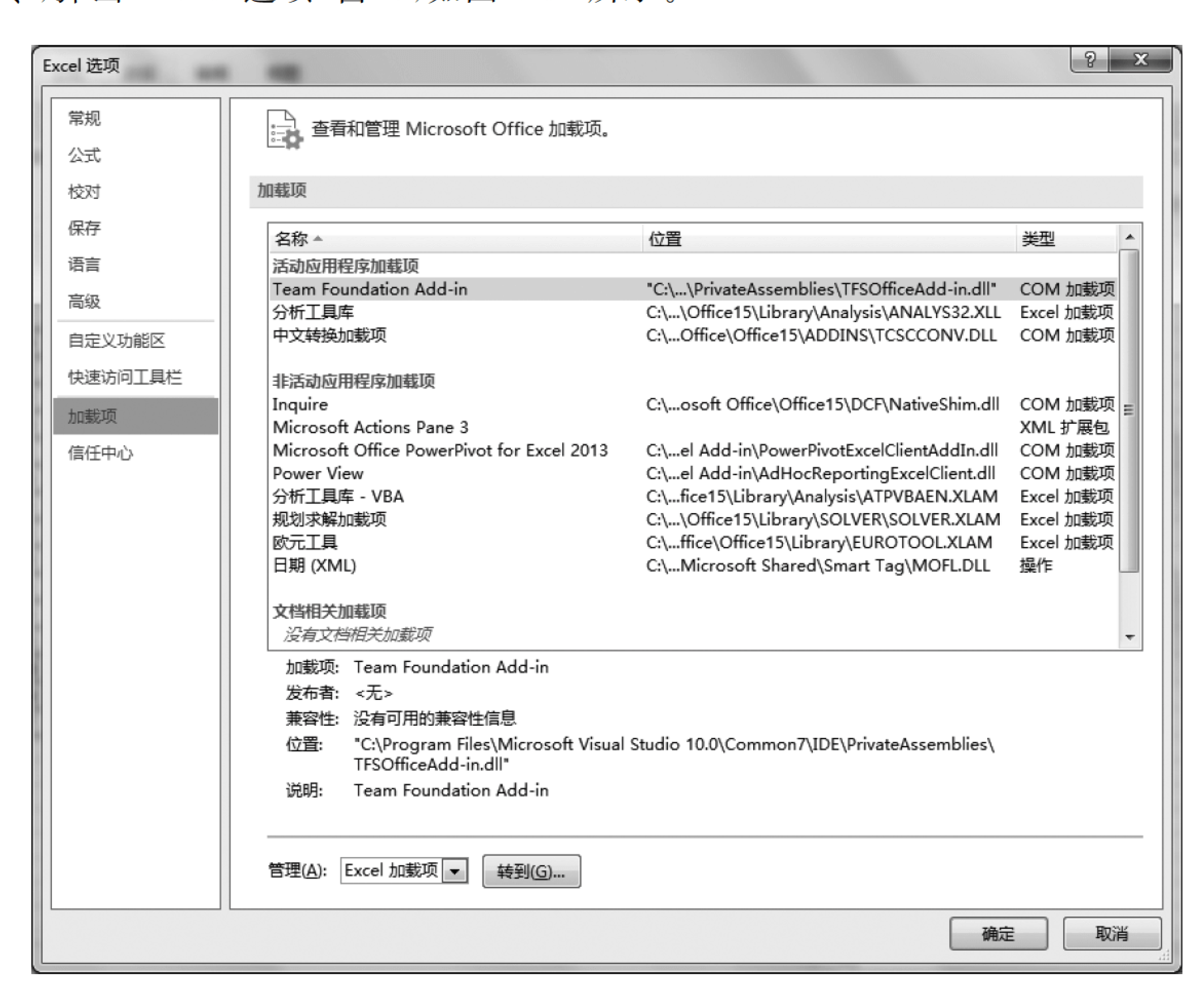


图 3-50 "Excel 选项"窗口

2. 在"Excel 选项"窗口中，先在左边的"常规"列表框中选择"加载项"，然后在下边单击"转到"按钮，弹出如图 3-51 所示的"加载宏"窗口。

3. 在"加载宏"窗口中，勾选"分析工具库"，单击"确定"按钮。

这时再单击"数据"菜单，在工具栏的最右边将出现"数据分析"按钮。



图 3-51 “加载宏”窗口

参考文献

[1] 刘红阁,王淑娟,温融冰. 人人都是数据分析师——Tableau 应用实战[M]. 北京:人民邮电出版社,2015.

[2] Tableau 白皮书

第 4 章 大数据分析的思维特征

4.1 大数据应用分析的实务框架

4.1.1 大数据应用分析的四个层面

大数据应用分析从实务角度可以划分为四个层面,如图 4-1 所示。

第一个层面:器。主要指分析数据用的利器,包括硬件和软件两大方面。硬件包括计算机、移动设备、传感器、视频音频设备等;软件包括数据库系统、文字处理软件、数据分析软件、数据采集软件、数据转换软件、图像处理软件和专用的工具模块等。

第二个层面:技。主要指分析数据的技术和方法,包括四个方面的主要内容。

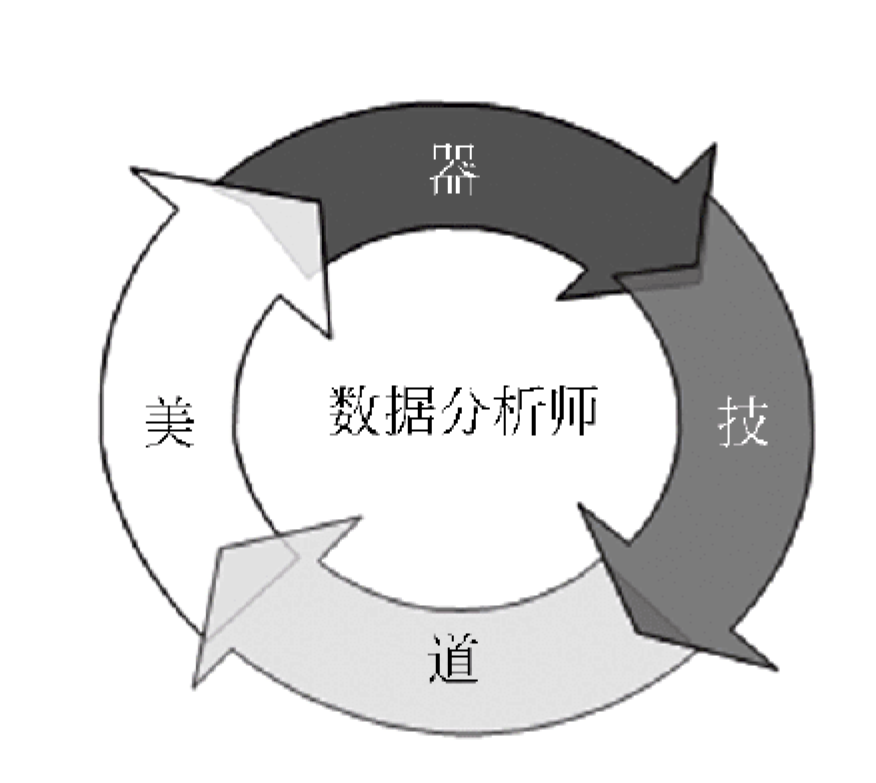


图 4-1 大数据应用分析的四个层面

- 方法。有适用于大数据的一些技术,包括大规模并行处理(MPP)数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网、可扩展的存储系统,以及具体的技术和方法,例如,如何在系统庞大的企业管理软件中下载数据,如何采集大型关系数据库中的数据,如何采集视频音频数据,如何在动态数据中定位采集分析需要的数据,当数据被恶意删除后如何恢复,如何利用卫星遥感图片(简称卫片)和航片数据,如何进行统计分析、趋势分析、回归分析、挖掘分析。再细一步讲在运用地理信息系统软件对空间数据开展分析时,如何交集取反、擦除、相交、裁剪、缓冲区分析、拓扑检查等。
- 参数。为了解读数据的含义,做出明确的判断,必须有对照的标准数据,标准数据包括法律法规、行业标准、技术参数、历史数据等。
- 函数。数据分析人员会在平时的分析中积累大量的模块,建立常用的函数库。
- 案例。在数据分析工作中经历过的具有典型性和普遍借鉴意义的事件总结。

在方法、参数、函数和案例四个方面中,方法、参数是公开的、共享的;函数常常带有私有性、专属性的特点,需要数据分析人员凭借自己的努力积累和沉淀;经典案例则具有更大的放射性效果,仁者见仁,智者见智,每个人都可以品味出有益的味道。

第三个层面:道。指分析数据的思维方式。大数据的数据量巨大、类型繁多、瞬息万变,如何在浩瀚无际的汪洋大海中捞到细如毫发的一根针?关键是要有一个清晰明确的分析思路,我们称之为大数据分析的思维方式。大数据分析的思维方式可以有多种,其中基础的是特征发现。特征发现包括特征枚举、特征捕捉和特征分析三个步骤。特征发现

的前提是假定任何行为都是有痕迹、有特征的。归纳不同行为的特征,然后去观察分析对象中有没有这类痕迹,若发现有类似特征,就捕捉含有此类特征的痕迹数据,进行解读分析。例如在地震前夕,会出现很多异常现象,鸡会乱飞、狗会狂叫、水位会变化、山体的移动会加剧等。2015 年 4 月 25 日 14 时 11 分,尼泊尔发生 8.1 级强烈地震。一位中国游客描述当时的情景说:“动物比人警觉,在有震感前,广场上一大群鸟突然全都飞了起来。原本懒散地趴在桥上的印度神牛也狂奔起来。”集中发生这种现象常常预示着可能发生地震,有了这种分析的前提,数据分析师就可在多个观测点中观察有没有这些情况发生,如有,可及时采集、整理相关的数据,然后进行解读,做出判断。在特征发现思维方式中包含三个要素:公理、数据、演绎。以在我国许多地方流传的“八月十五云遮月,正月十五雪打灯”这句气象谚语为例,这是千百年来古人观察和经验的结晶,大家都认为是正确的,是不容置疑的。这就是“公理”。然后人们观察农历八月十五这一天是否出现了云遮月,是或不是都记录下来(采集),这就是数据。最后得出结论,如果是,则正月十五要雪打灯,否则就不打。这是演绎推理做出判断。

第四个层面:美。这里的“美”是指审美活动。为什么要谈起这个问题呢?在上面一段论述中我们谈到,特征发现思维方式包括公理、数据、演绎三个要素。公理是先人们发现总结并且被人们所公认的,奉为圭臬,是分析数据的标准。有了这个标准,我们才能开展分析活动,才能解析数据的意义。但是公理、定律这类东西是被逐步发现的。世界越发展、人类的研究活动越深入,数据的联系也越来越复杂,需要研究和回答的困惑或问题就越多,出现了许多新变化、新情况、新问题,原来的公理、定律可能不够用了,有的可能还暴露出了缺陷和问题,需要完善或者推翻重来。如何在没有公理、定律这些标准的时候去开展分析呢?如何在前人从未遇到过的数据面前开展数据分析呢?这个时候特别需要强调直觉和想象力。很多科学家都认为,在科学研究中要想有所发现和发明,要想获得创造性的成果,必须依赖直觉和想象。爱因斯坦十分强调想象、直觉、灵感在科学研究中的作用。他认为,科学体系中的概念和命题都是思维的自由创造,所以必须突破形式逻辑的局限。他说:“我相信直觉和灵感。”“想象力比知识更重要,因为知识是有限的,而想象力概括着世界上的一切,推动着进步,并且是知识进化的源泉。严格地说,想象力是科学研究中的实在因素。”他还说:“物理学家的最高使命是要得到那些最普遍的基本规律,由此世界体系就能用单纯的演绎法建立起来。要通向这些定律,并没有逻辑的道路,只有通过那种以对经验的共鸣的理解为依据的直觉,才能得到这些定律。”

叶朗先生在《美学原理》中把依赖直觉和想象力的研究活动归纳在科学美的范畴。科学美主要是一种数学美、形式美。杨振宁认为,理论物理学存在三种美:现象之美、理论描述之美、理论架构之美。现象之美是指物理现象之美。理论描述之美是指一些物理定律有一种很美的理论描述,如热力学的第一、第二定律就是对自然界的某些基本性质的很美的理论描述。理论架构之美是指一个物理学的定律公式化时,它趋向一个美的数学架构。这种物理学的理论架构,以“极度浓缩的数学语言写出了物理世界的基本结构”,是一种深层的美。追求科学美是科学研究的一种动力,很多科学家都相信对美的追求可以把我们引向真理的发现。很多有原创性的物理学家都说,他们的创见是在灵感的一闪中获得的,不是一点一滴地推敲,也不是按逻辑过程进行分析推理,而是突然间有如神助地出

现了。大数据本身是十分枯燥和冰冷的，数据分析师如果能把烦琐的分析变成一种发现规律的审美活动，相信分析过程会充满乐趣和奇迹，充满审美过程的享受。[①]

4.1.2 四个层面的关系

如图 4-2 所示，上面介绍的器、技、道和美四个层面可对应到大数据分析的四个方面，分别是：器——工具软件；技——技术方法；道——思维方式；美——感觉想象。

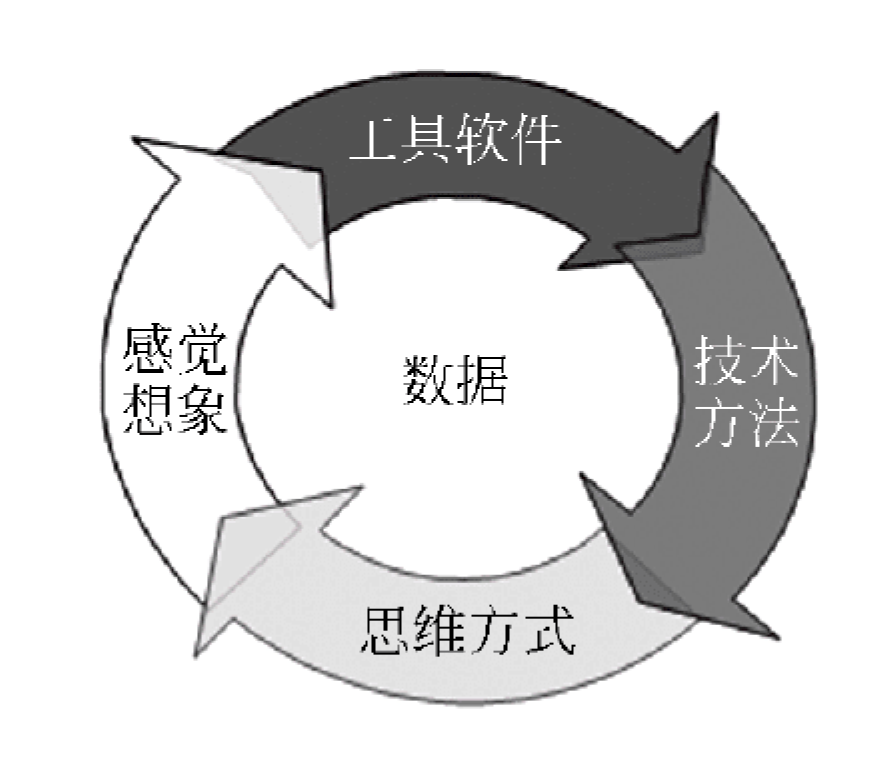


图 4-2 四个层面的相互关系

在大数据分析的完整过程中，数据是核心，是分析的对象。围绕数据中心，器是工具，是分析师手中的武器。大数据是数字化的，肉眼是不可见的，而且有些数据需要在一定的语境下才能识读，如传感器数据、卫星数据等，要经过多次的转换翻译，才能辨识。同时，所有数字化数据都是有严格格式的，都需要在特定的系统中才能处理。更为重要的是，在许多大数据分析软件和工具中开发厂商投入了艰辛的研究，提供了许多成熟的方法和技术，固化了许多经验，为数据分析师提供了许多方便，非常有益于开展分析。器是基本、是前提、是必不可少的。但是有了这个工具，如何使用，使之发挥最大的功能，就来到了第二个层面——技术和方法。工具再好，即便是具有学习能力的软件也是人设计的，要发挥它的作用，需要使用者掌握熟练的技能和技巧。对一些经典技术和方法要反复训练、反复实践，达到熟能生巧的境界。在掌握了熟练的技术以后，能否有效地开展分析，从数据的矿藏中开发和冶炼出真金白银就要看分析者的思维方式是否科学、概括特征是否准确、捕捉痕迹是否敏锐、辨析规律是否科学。在全部的分析过程中，思维方式是一个纲，是思路、是灵魂，是统领全部分析活动的。分析思路是否正确、思维是否清晰，常常决定着分析活动的成败高下。事情到这里并没有结束。有时候分析活动会出现这样的现象，分析师做了大量工作，但分析始终停留在一定的水平上，挖掘不出价值更大的宝藏。这就要从美的层面上去寻找原因，最根本的原因是分析者缺乏创造性思维，尤其是遇到从来没有见到过的数据、面对从来没有开展过的分析时，面临缺乏标准的困惑时缺少感觉和想象力，缺少灵光闪现。

从上面的简要论述中，我们也可以体会到，在数据分析动态过程中，器、技、道和美是循环往复出现的，各司其职，完美融合在一起发挥作用，相互补充，相得益彰。

4.2 大数据分析的特征发现

大数据中的数据量巨大，类型繁多，来源多处，真可谓千头万绪、盘根错节，金子常常被掩埋在厚厚的沙堆之中，要对这样的数据展开分析，运用功能强大的分析工具，掌握分析的技术和方法，都是十分必要的。但仅仅有这些还是不够的，核心的问题是分析师必须有清晰的分析思路，培养科学的思维方式，这就是特征发现。本节将结合具体分析案例，

① 以上爱因斯坦的几段话和杨振宁先生的论述转引自叶朗. 美学原理[M]. 北京：北京大学出版社，2009：第七章.

进一步讨论这种思维方式。

4.2.1 特征发现的案例

2015年4月25日星期六凤凰网发表了一则新闻。

原标题：他让股市5分钟蒸发近万亿美元

英国期货交易员纳温德·辛格·萨劳21日因被美国司法部指控涉嫌操纵市场而被英国警方逮捕，面临引渡。一桩2010年美国股市"闪电崩盘"的陈年旧案由此再次被媒体翻开，而这名交易员则成为争议焦点。

是谁，那一天让纽约股市道琼斯指数在5分钟内暴跌600点，总市值蒸发近1万亿美元，而自己牟利近百万美元？英国媒体23日刊登了萨劳的照片，让外界首次一睹这名"华尔街之狼"的真容。

2010年5月6日，道琼斯指数在20多分钟内暴跌约1 000点，其中最剧烈的600点下跌发生在5分钟内，之后指数又大幅回升。这一交易日也创下美国股市有史以来最大单日盘中跌幅，堪称华尔街历史上波动最为剧烈的20分钟。

美国执法部门调查发现，事件的罪魁祸首是来自英国的萨劳。他当时在位于伦敦郊区的普通住宅内，利用家用电脑对美国股市的期货交易系统下虚假合约单，制造恐慌并引发股市动荡。

美方指控，萨劳利用一个计算机交易程序对美股股指期货下巨额卖单，但能瞬间实现撤单，以保证这些卖单不会成交，却能对交易价格构成实时抛压。这一做法的目的并非完成交易，而是影响价格和达到操纵市场的目的，因而构成欺诈。

5月6日那一天，萨劳自上午开始就对交易系统下虚假卖单，市场出现下跌趋势后，他继续加大"抛售"力度，在中午12时33分左右达到最疯狂阶段，致使美股指数暴跌。萨劳随后在暴跌的地点购进数只"便宜"期货合约，待股指回升后抛售，当天盈利近90万美元。

美国司法部估计，2010—2014年，萨劳通过交易美股标普500指数期货合约总共盈利4 000万美元。

萨劳如今面临美方司法部门提起的一项电信欺诈、10项大宗商品欺诈、10项大宗商品市场操纵行为以及一项欺骗行为指控。如果被裁定成立，这些指控合计将为萨劳带来最高380年监禁。

无独有偶，2015年11月1日，新华社发布消息：上海公安机关成功侦破一起以贸易公司为掩护，境外遥控指挥、境内实施交易，作案手段隐蔽、非法获利巨大的涉嫌操纵期货市场的案件。2015年六七月间，中国证券期货市场出现异常巨幅波动，广大投资者蒙受巨大损失。针对相关部门移交和公安机关侦查掌握的一些违法犯罪线索，公安机关掌握了外商投资的伊世顿公司涉嫌操纵期货市场等犯罪的线索，遂交由上海市公安局依法开展立案侦查。专案组查明，伊世顿公司系外籍人员Georgy Zarya(音译扎亚)、Anton Murashov(音译安东)在香港各自注册成立一家公司后，于2012年9月用两家香港公司的名义在江苏省张家港保税区以美元出资注册成立的贸易公司。扎亚为伊世顿公司法定

代表人，安东负责技术管理。两人在公司成立前分别供职于欧洲的投资银行和期货公司，从事证券期货交易工作。受扎亚、安东指使，中国境内的犯罪分子为规避中国金融期货交易所相关规定的限制，先后向亲友借来个人或特殊法人期货账户 31 个，供伊世顿公司组成账户组进行交易。伊世顿公司以贸易公司为名，隐瞒实际控制的期货账户数量，以 50 万美元注册资本金以及他人出借的 360 万元人民币作为初始资金，在中国参与股指期货交易。安东及其境外技术团队设计研发出一套高频程序化交易软件，远程植入伊世顿公司托管在中国金融期货交易所的服务器，以此操控、管理伊世顿账户组的交易行为。伊世顿账户组通过高频程序化交易软件自动批量下单、快速下单，申报价格明显偏离市场最新价格，实现包括自买自卖（成交量达 8 110 手、113 亿元人民币）在内的大量交易，利用保证金杠杆比例等交易规则，以较少的资金投入反复开仓、平仓，使盈利在短期内快速放大，非法获利高达 20 多亿元人民币。2015 年 6 月初至 7 月初，证券期货市场大幅波动，伊世顿公司在交易沪深 300、中证 500、上证 50 等股指期货合约过程中，卖出开仓、买入开仓量在全市场中位居前列，该公司账户组平均下单速度达每 0.03 秒一笔，一秒内最多下单 31 笔，且成交价格与市场行情的偏离度显著高于其他程序化交易者。以 6 月 26 日的中证 500 主力合约为例，该公司账户组的卖开量占市场总卖出量 30% 以上的次数达 400 余次；以秒为单位计算，伊世顿账户组的卖开成交量在全市场中位列第一的次数为 1 200 余次；其卖开成交价格与市场行情的偏离度为当日程序化交易者前 5 名平均值的 2 倍多。据统计，仅 6 月初至 7 月初，该公司账户组净盈利就达 5 亿余元人民币。监管机构认为，伊世顿公司的期货交易行为扩大了日内交易价格波幅，与市场价格走势存在关联性，影响了当时的市场交易价格和正常交易秩序。公安机关认为，伊世顿公司的异常交易行为符合操纵股指期货市场的特征，涉嫌操纵期货市场犯罪。侦查还表明，伊世顿公司将巨额非法获利中的近 2 亿元人民币通过“地下钱庄”转移出境，交给安东等境外人员。

读了上面的新闻，相信很多读者对股市和期货市场的风险会有更进一步的理解，同时也会更加深切地认识到打击内幕交易、操纵市场等违法行为的必要性。其实，这也是监管部门努力追求的一个目标。而要达到这个目标，大数据分析是锐利的武器，也是必不可少的。下面是一个内幕交易案件的完整查处过程。

案件查处背景

2007 年我国的 A 股一路上扬，上证指数到 10 月 16 日达到了 6 124 点，从 10 月下旬开始震荡，然后下跌，到 2008 年 10 月 28 日下跌到 1 664 点，坐了一个大大的过山车。2008 年夏季当股市一路下跌的时候，一个检查组进入一家证券公司，检查股市交易中存在的问题。当时面临的形势是股市还在不停地震荡，许多股民，尤其是许多散户被深度套牢，笼罩在一片阴云之中。股市怎么了？有没有内幕交易？有没有操纵市场？有没有老鼠仓？怎样才能有效地监管股市？怎样才能保护广大股民的合法权益？这些都成了从上到下，从管理层到普通股民一致关心的问题。股市风谲云诡，数据海量，进出频繁，头绪杂乱，从哪里入手，怎样才能发现其中的问题呢？难度之大，难以想象。检查组面临严峻的考验。怎么办？检查人员反复讨论，决定先召开座谈会。

集思广益确定检查方向：资金内转

检查组组织了由证券公司管理层、操盘手、股评分析师等参加的多个座谈会，广泛听取

大家对开展证券市场检查的建议。功夫不负有心人，在汇总讨论意见的时候，大家发现尽管各个组参加的对象不同，看问题的角度不同，但所有的讨论都提出了一个建议，要重点开展对资金内转账户的检查。什么是内转账户呢？就是有的账户开了户后不是买卖股票，而是大量频繁地转移资金，常常是多个账户之间相互划转，转来转去，有的最后又转回到源账户画了一个圆，有的资金被转走，不见了踪影。这样频繁地转移，本身就掩盖着一定的目的，这类账户应当列为检查关注的重点。大家统一了认识，确定检查方向是：资金内转。

运用特征分析的思维方式展开检查

在实施检查中，检查组运用特征分析的思维方式，有条不紊，层层推进，精准延伸。

第一步：特征枚举

大家查阅法律法规，上网寻找国内外的案例，借鉴司法部门的案件，列举资金内转的所有表现形式，了解资金内转是如何操作的，列举出了多种方式。如：

- 将大额资金分散转至若干账户；
- 资金转入后立即购买股票；
- 多个账户向一个账户转入资金；
- 相关联的内转账户集中购买一只股票；
- 相关联的内转账户集中卖出一只股票；

……

通过特征枚举，检查人员加深了对资金内转操作方式的了解，为进一步查处奠定了基础。

第二步：特征捕捉

首先，采集整理数据，构建审计中间表。从该证券公司北京、上海、武汉、深圳等各个节点采集相关数据后，检查人员经过筛选整理，生成了客户交易流水中间表及客户信息中间表两张数据表。其中客户交易流水中间表（如图4-3所示）记录该证券公司业务客户的A股交易流水明细情况；客户信息中间表（如图4-4所示）记录客户基本信息明细情况。在对客户交易流水数据的浏览观察中，检查人员发现了一个数据的重要属性，凡是内转交易，“摘要代码”字段值均是代码“140025”。

其次，查询有资金内转的账户。

查询1：筛选出有资金内转的账户

```
select  客户代码,客户姓名,资金账号,sum(资金转出) 转出合计
into  资金转出大户  from  主表_北京2_A股交易流水表
where 摘要代码='140025'
group by  客户代码,客户姓名,资金账号
order by  转出合计
```

对“资金转出大户”表的查询结果如图4-5所示。

分析“资金转出大户”表，发现有一个叫孙××的客户将其资金账户内的171万元分别转至苗××等12个自然人的资金账户中。

查询和分析孙××的资金转出记录。

查询2：筛选出“孙××”的资金转出记录

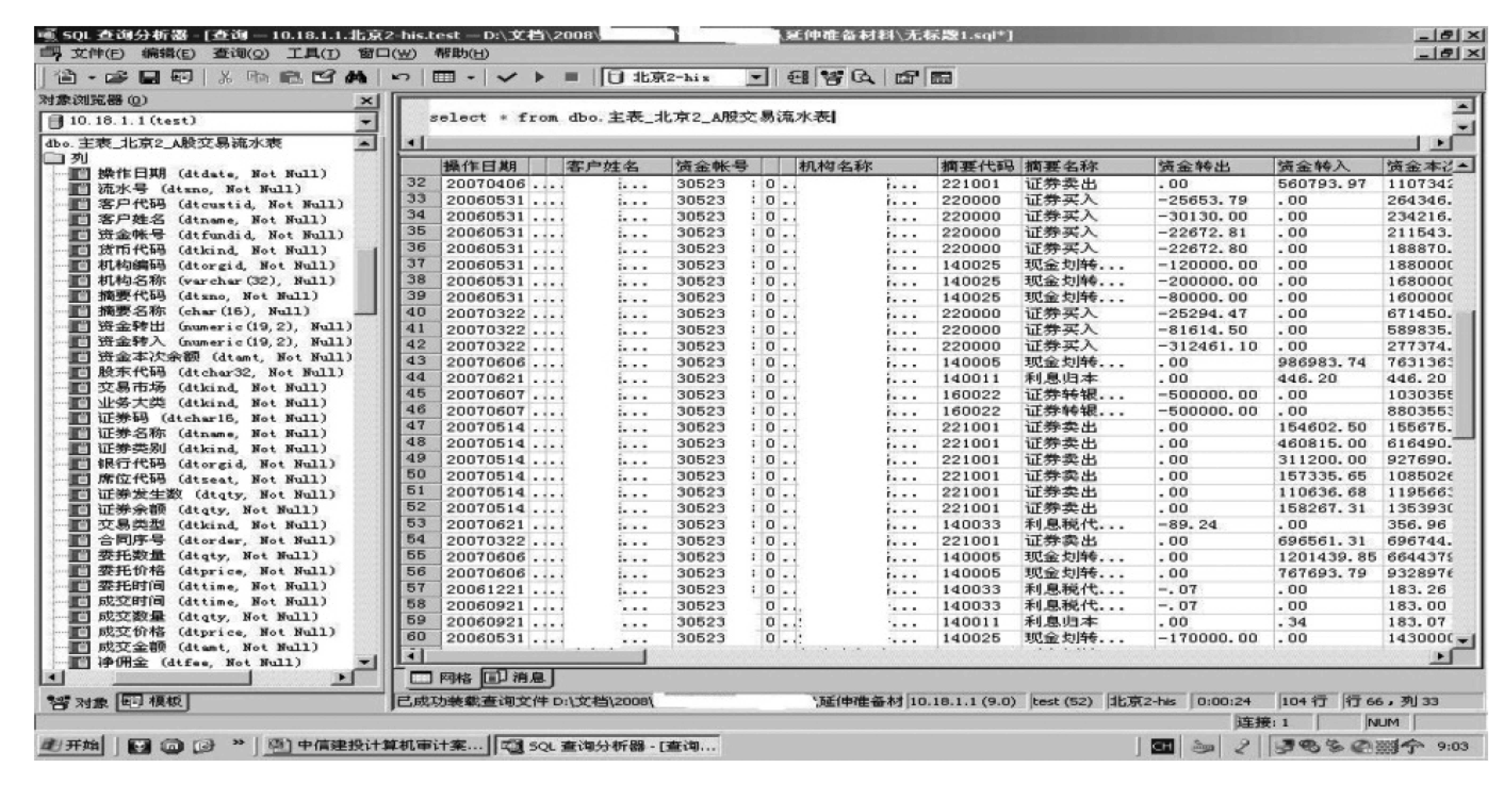


图 4-3　客户交易流水中间表

图 4-4　客户信息中间表

```
select * into 孙××转出
from 主表_A股交易流水表
where 摘要代码 ='140025' and 客户代码='XXXXX0023613'
```

Microsoft Excel - 资金转出大户.XLS

序号	客户代码	客户姓名	资金帐号	转出合计
1	38839		92809880	-10793177.8
2	01022		90900178	-10000000
3	74521		39309986	-8826000
4	38383		92809416	-4592900
5	22587		98242889	-3308320.62
6	38189		92809213	-2890000
7	13278		90713278	-2887800
8	20396		90720395	-2807789
9	38776		92809820	-2681500
10	21025		98241248	-2660000
11	20090		90720090	-2200000
12	20123		90720123	-1935800
13	00842		90200953	-1827216.55
14	18763		90718763	-1814334.07
15	10712		30621738	-1800000
16	23613		30523758	-1710000
17	19689		90719685	-1456429.6
18	29252		90723565	-1400000
19	24949		30630455	-1307417.48

祁楠_资金转出大户

图 4-5　有资金内转的全部记录

该语句筛选出的数据如图 4-6 所示。

Microsoft Excel - 查询3_[redacted]资金转出.XLS

操作日期	客户代码	客户姓名	资金帐号	机构名称	摘要代码	摘要名称	资金转出	资金本次余额
20060531	023613		23758	路营业部	140025	现金划转取	-120000	1880000
20060531	023613		23758	路营业部	140025	现金划转取	-200000	1680000
20060531	023613		23758	路营业部	140025	现金划转取	-80000	1600000
20060531	023613		23758	路营业部	140025	现金划转取	-170000	1430000
20060531	023613		23758	路营业部	140025	现金划转取	-110000	1320000
20060531	023613		23758	路营业部	140025	现金划转取	-120000	1200000
20060531	023613		23758	路营业部	140025	现金划转取	-130000	1070000
20060531	023613		23758	路营业部	140025	现金划转取	-150000	920000
20060531	023613		23758	路营业部	140025	现金划转取	-100000	820000
20060531	023613		23758	路营业部	140025	现金划转取	-150000	670000
20060531	023613		23758	路营业部	140025	现金划转取	-180000	490000
20060531	023613		23758	路营业部	140025	现金划转取	-200000	290000

祁楠_孙　转出

图 4-6　发现孙××有资金内转行为

查询 3：筛选出与“孙××”转出记录对应的资金转入记录

```
select * into 12人转入
from  主表_A股交易流水表
where 操作日期='2006-5-31' and 摘要代码='1400025'
   and 资金转入 in(120000,200000,110000,130000,80000,
                       170000,150000,100000,180000)
```

查询 4：筛选出“孙××”等 13 户的全部交易流水

```
select * into 孙××全部
from  主表_A股交易流水表
where  客户代码='×××××0023613'
```

查询 5：筛选出“孙××”等 13 户关联内转账户的开户情况

```
select * into  13户开户资料
from  客户基本信息表
where 客户代码 between '×××××0023618' and '×××××0023631'
    or 客户代码='×××××0023613'
```

该语句筛选出的数据如图 4-7 所示。

操作日期	客户代码	客户姓名	资金帐号	机构名称	摘要名称	资金转入	资金本次余额
20060531	23619		23743	路营业部	现金划转存	200000	200000
20060531	23620		23745	路营业部	现金划转存	180000	180000
20060531	23621		23746	路营业部	现金划转存	150000	150000
20060531	23622		23747	路营业部	现金划转存	100000	100000
20060531	23623		23748	路营业部	现金划转存	150000	150000
20060531	23624		23749	路营业部	现金划转存	130000	130000
20060531	23625		23750	路营业部	现金划转存	120000	120000
20060531	23626		23751	路营业部	现金划转存	110000	110000
20060531	23627		23752	路营业部	现金划转存	170000	170000
20060531	23628		23753	路营业部	现金划转存	80000	80000
20060531	23629		23755	路营业部	现金划转存	200000	200000
20060531	23630		23756	路营业部	现金划转存	120000	120000

图 4-7　发现与孙××交易的账户客户代码连续

查询 6：筛选出 12 户关联内转账户股票卖出后将资金转回“孙××”资金账户的记录

```
select * into 孙××转入
from  主表_A股交易流水表
```

where 摘要代码='1400025' and 客户代码='×××××0023613'

该语句筛选出的数据如图 4-8 所示。

客户代码	客户全名	联系地址	邮政编码	电话号码	开户日期
023613	孙	市芝罘区文化宫后街88号	4000	054555556	20060314
023619	苗	大海阳145号	4000	65642	20060322
023620	邹	大海阳145号	4000	65642	20060322
023621	孙	西南关南街3-7号	4000	65642	20060322
023622	丁	大海阳145号	4000	65642	20060322
023623	丁	大海阳145号	4000	65642	20060322
023624	张	黄集乡西张庄新镇村	4000	65642	20060322
023625	刘	大海阳145号	4000	65642	20060322
023626	邹	平里店镇石柱兰村515号	4000	65642	20060322
023627	刘	大海阳145号	4000	65642	20060322
023628	黄	王家庄镇孟家官庄村	4000	65642	20060322
023629	徐	大海阳145号	4000	65642	20060322
023630	韩	生辰街道梨园村62号	4000	65642	20060322

图 4-8　发现与孙××交易的联系人电话号码相同

进一步分析发现具体过程如下：2006 年 5 月，孙××通过银证转账的方式分四笔将 200 万元转入其在某证券公司营业部的资金账户，并将其中的 29 万元用于购入“＊ST××”股票，与此同时他又将剩余的 171 万元分别转入苗××等 12 个自然人的资金账户，并全部用于购入“＊ST××”股票。2007 年 3 月至 5 月，上述 13 个账户中的“＊ST××”股票被陆续卖出，在不到一年的时间里合计盈利 861.86 万元，收益率高达 430%。2007 年 6 月，苗××等 12 人的资金账户中的资金被全部转回孙××的资金账户，孙××资金账户的全部资金于第二日即被以银证转账的方式转出。从 2006 年 5 月 200 万元资金转入至 2007 年 6 月全部资金转出，这 13 个账户基本无其他股票买卖行为发生。通过查询“客户信息中间表”中上述 13 个账户的开户资料还发现，除孙××外，其余 12 户均为 2006 年 3 月同一天开户，客户代码连续，所留联系电话及邮政编码等信息完全相同。通过延伸孙××的开户行发现，在 2006 年 5 月孙××将 200 万元转入其证券公司资金账户的前一天，一个名为徐××的人将其个人结算账户内的 200 万元转入了孙××的账户，而此前孙××的账户内的余额仅为 1 元钱。最后发现，徐××正是“＊ST××”的收购重组方——某实业集团有限公司的董事长。

第三步：特征分析

对照我国相关的证券法律法规，徐××是“＊ST××”的收购重组方——某实业集团有限公司的董事长，是内幕人。他掌握了证券交易活动中涉及公司的经营、财务或者对公

司证券的市场价格有重大影响的尚未公开的信息，而且在内幕信息的价格敏感期内买卖相关股票，组织他人买卖相关股票，泄漏该信息。徐××的行为属于典型的内幕交易。

检查组上报的检查结果，引起了管理层的高度重视，对涉案人员进行了严肃处理。

分析了上面的案例，再来讨论特征发现的思维方式，会有更真切的感受。

4.2.2　特征发现的概念

特征是指可以反映事物特点的征象、标志等。特征发现实际上是指从大数据中提取有用的信息和知识的过程。

大数据的特征发现可以分为已知事件的特征发现、未知事件的特征发现及征兆发现等。已知事件的特征发现是指数据分析人员主要依据历史案例、业务处理逻辑等建立模型进行特征发现。在分析过程中，通常已知某些行为的特征表现，列举出特征，然后运用一定的技术方法寻找符合特征的数据，并进一步分析取证。未知事件的特征发现是指运用数据挖掘等技术方法发现事件的特征，这些特征在得出挖掘结果之前分析人员是无法预测的。而征兆发现与一般特征的发现有很大的差异，特征是指事件（问题、案件）已经发生，而征兆则是指事件尚未发生或正在进行当中。因此，对已知事件、未知事件的特征发现及征兆发现的一般过程和技术方法都是不同的。为了表述和理解的方便，本书在讨论特征发现时是作为一个大的概念使用的，包含了已知事件的特征发现、未知事件的特征发现及征兆发现三种情况。

在物联网、互联网、云计算、卫星跟踪定位等日益发达的今天，任何行为，包括人的行为、大自然的行为、社会的行为、经济的行为、机器的行为等都会实时留下痕迹。这些痕迹，有行为痕迹，行为痕迹记录活动的过程；有系统痕迹，系统痕迹是在计算机处理为基础的网络系统中留下的印记，如系统日志文件的数据；有数据痕迹，数据痕迹是在数据库和其他数据记录、处理、存储介质中留下的记载。这些痕迹中具有代表性的，能够表现其特点的被称为特征。这三种类型的特征互相联系、互相映射。行为特征映射系统特征、数据特征，系统特征映射行为特征、数据特征，数据特征映射行为特征、系统特征。这种联系和映射是特征发现得以实施的客观前提和基础。

4.3　对数据的分类

在大数据环境下，数据类型十分复杂，有来自天上的，如航天、卫星数据，有来自地下的，如地震观测数据；有来自人的，如管理数据、财务数据、文本数据，有来自机器的，如物联网传感器数据；有事后的数据，如财务报表；有实时的数据，如录像监控数据……在对数据展开分析时，可以按需要从不同的角度进行分类。采取何种标准、如何划分类型，要服务于分析的需要。如果要分析数据的变化，可以用时点划分数据；如果要强调数据的出处，可以用数据的来源来划分；如果要强调数据的格式，可以用是否具有典型的结构来划分，等等。

在数据的应用分析中，有一种分类经常被分析师采用：把数据分成数值型数据和非数值型数据两大类。数值型数据包括数值类型、货币类型、日期类型和字符串类型的数

据。这是我们在分析实践中遇到最多的情况。非数值型数据如文本文件、图像、声音乃至网页、社交网站、传感器等其他类型的数据。在分析实务中,这两类数据常常融合在一起使用。非数值型数据常常激发分析人员的灵感,帮助形成分析思路。数值型数据常常作为查询、多维、挖掘分析的对象,从中发现规律,锁定证据。

为了更好地说明两类数据的结合在数据分析中的作用,我们再看一个土地整治审计分析的案例。

土地整治是盘活存量土地、强化节约集约用地、适时补充耕地和提升土地产能的重要手段,是保障发展、保护耕地、统筹城乡土地配置的重大战略。土地整治项目有三个特点:一是项目面积大,一个省辖市一年内验收确认的土地开发整理项目面积往往达几十万亩,单个项目动辄上千亩,即便是实地测量项目区面积都很困难,更谈不上对项目区内耕地、园地等明细地类进行深入分析;二是项目分布散,中央、省、市、县和乡镇各级政府都有土地整治项目的投入,项目数量多,由于资金分配会考虑区域平衡,项目会遍布各个区县、各个乡镇甚至各个行政村,单个项目平均投资比较小,项目和资金都非常分散;三是地形复杂、交通不便,很多项目分布在沟垅里、山头上,现场很难到达,项目查看效率非常低。在这种情况下,依靠丈量、观察和计算等传统的检查方法,根本无法实现审计目标。在地理信息系统环境下,利用 ArcGis 和谷歌地球等软件,检查人员可以有效克服上述困难,通过对项目区域、合适时点的土地利用现状数据、合适时点的遥感影像等数据进行分析,并借助外部数据,发现和分析在土地整治项目申报及管理中存在的问题。

一个审计项目组对 2008—2013 年某地所有土地开发项目每一个地块整理前后的实际状况进行了比对。审计人员把审计分析需要的各类数据,如数据库数据、Excel 数据、卫星遥感数据有机融合在一起分析,逐一比对,发现了该地区土地开发项目中存在的问题。

一、分析方法

1. 数据准备

将先前为检查目的采集的土地整治数据库附加到 SQL Server 中,然后将分市和分区县的土地整治数据表合并为一张数据表。

2. 通过 ArcGis 连接数据库

(1) 打开 ArcGis,并连接 SQL Server 数据库,将土地整治数据图斑导入 ArcGis 中,如图 4-9 所示。

(2) 转换坐标系。打开土地整治图斑,由于先前采集的土地整治图斑使用 Xian 1980 坐标系,谷歌地球影像图使用 WGS1984 坐标系,为了使土地整治图斑和影像图完美叠加需要调整坐标系。经配准计算,Xian1980 与 WGS1984 经度偏移 120 米,纬度偏移 50 米,以此为基准进行坐标转换,如图 4-10 所示。

(3) 筛选土地开发整理项目,分区县分项目导出图层。调整坐标系后,分区县分项目将土地整治开发整理项目按属性选择生成新的图层,具体操作如图 4-11 所示。

3. 导入 Google Earth,逐项目逐地块核实

(1) 由 ArcGis 生成 Google Earth 可识别的 KMZ 图层。将土地整治项目图斑转为 Google Earth 可识别的 KMZ 文件,具体操作如图 4-12 所示。选择“工具箱”→“转换工具”→



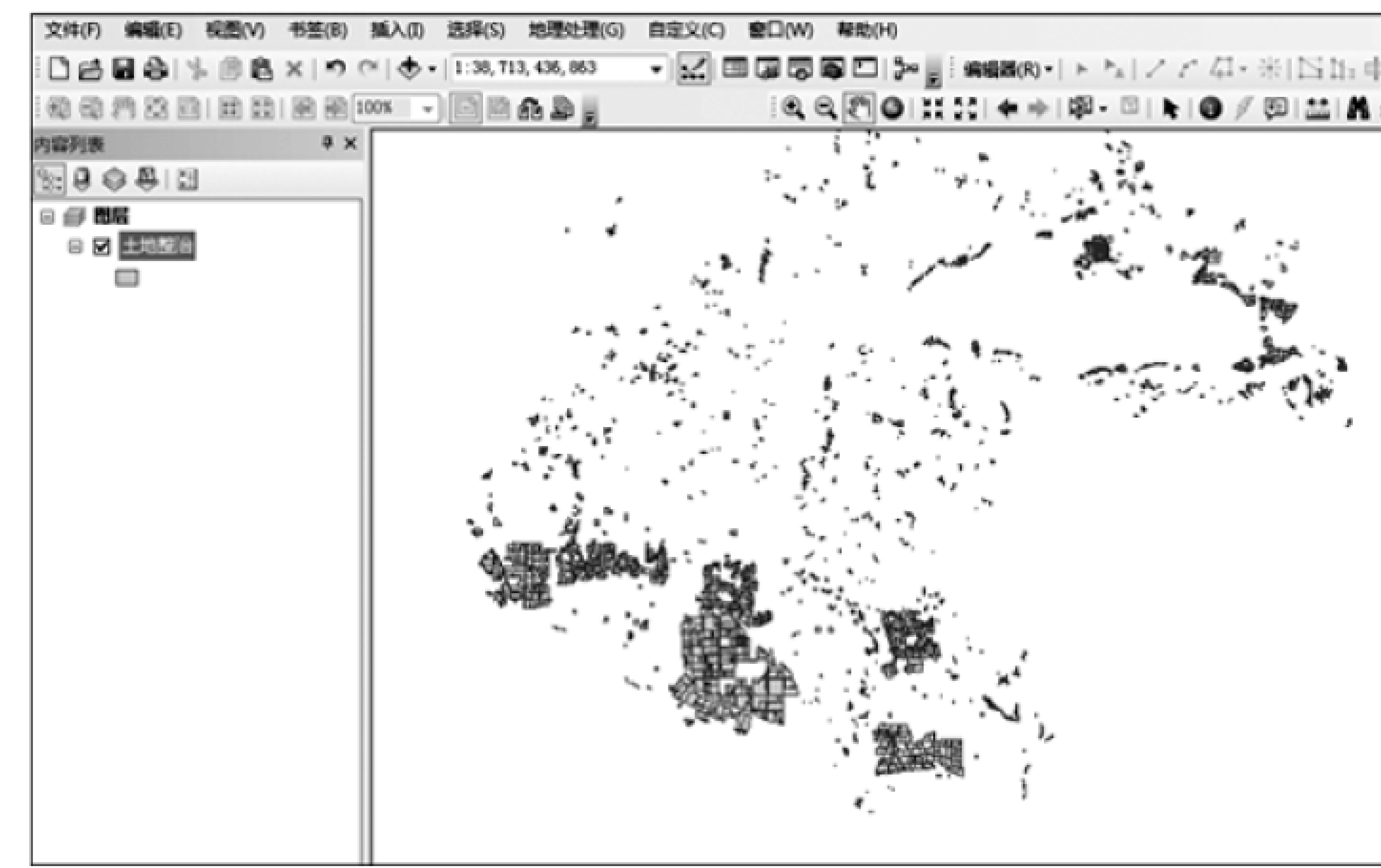


图 4-9　将数据导入软件中

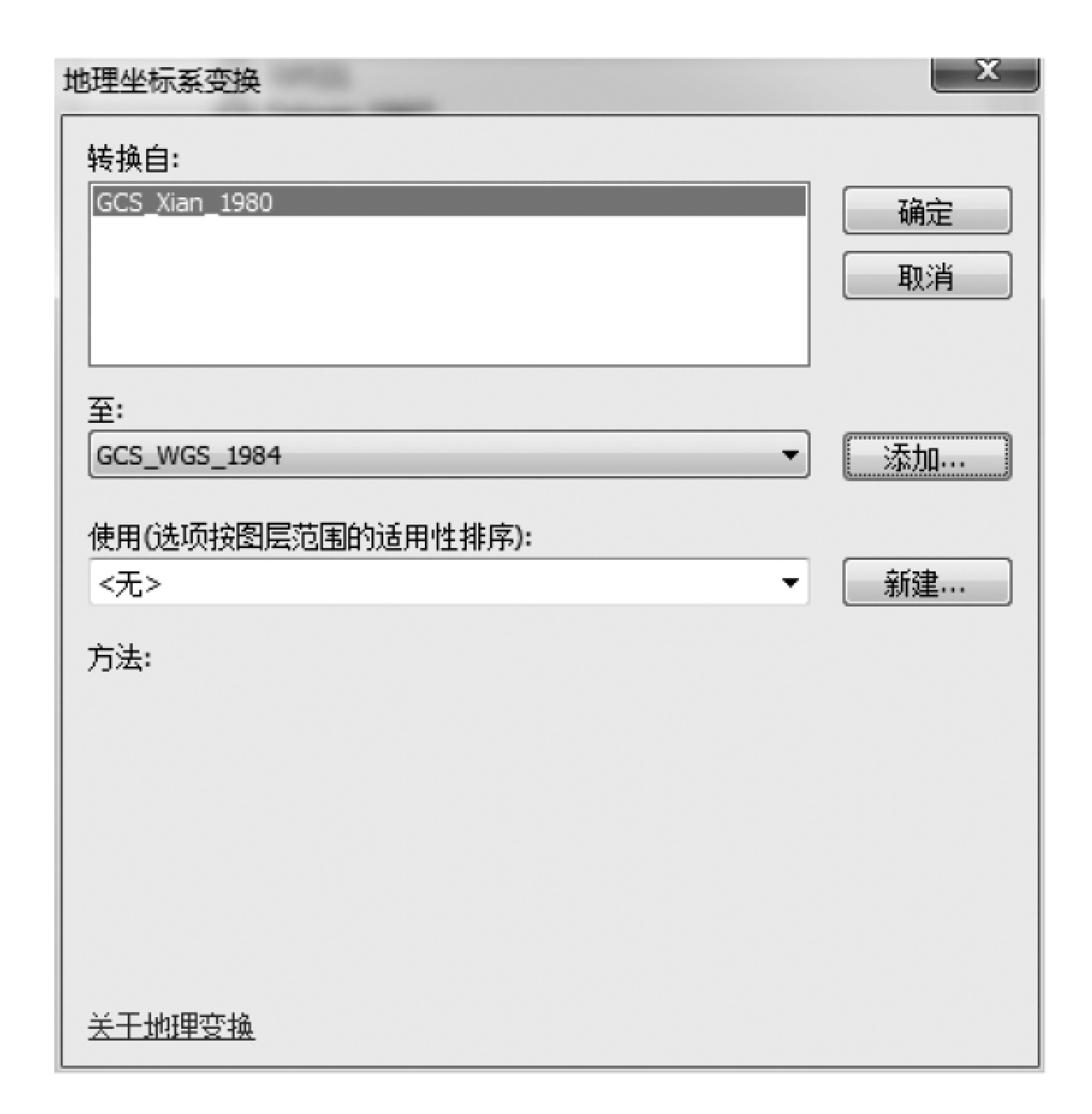




图 4-10　转换坐标





图 4-11　选择生成新图层

"转为 KML"→"图层转 KML",并选择该图层后单击"确定"按钮进行转换。

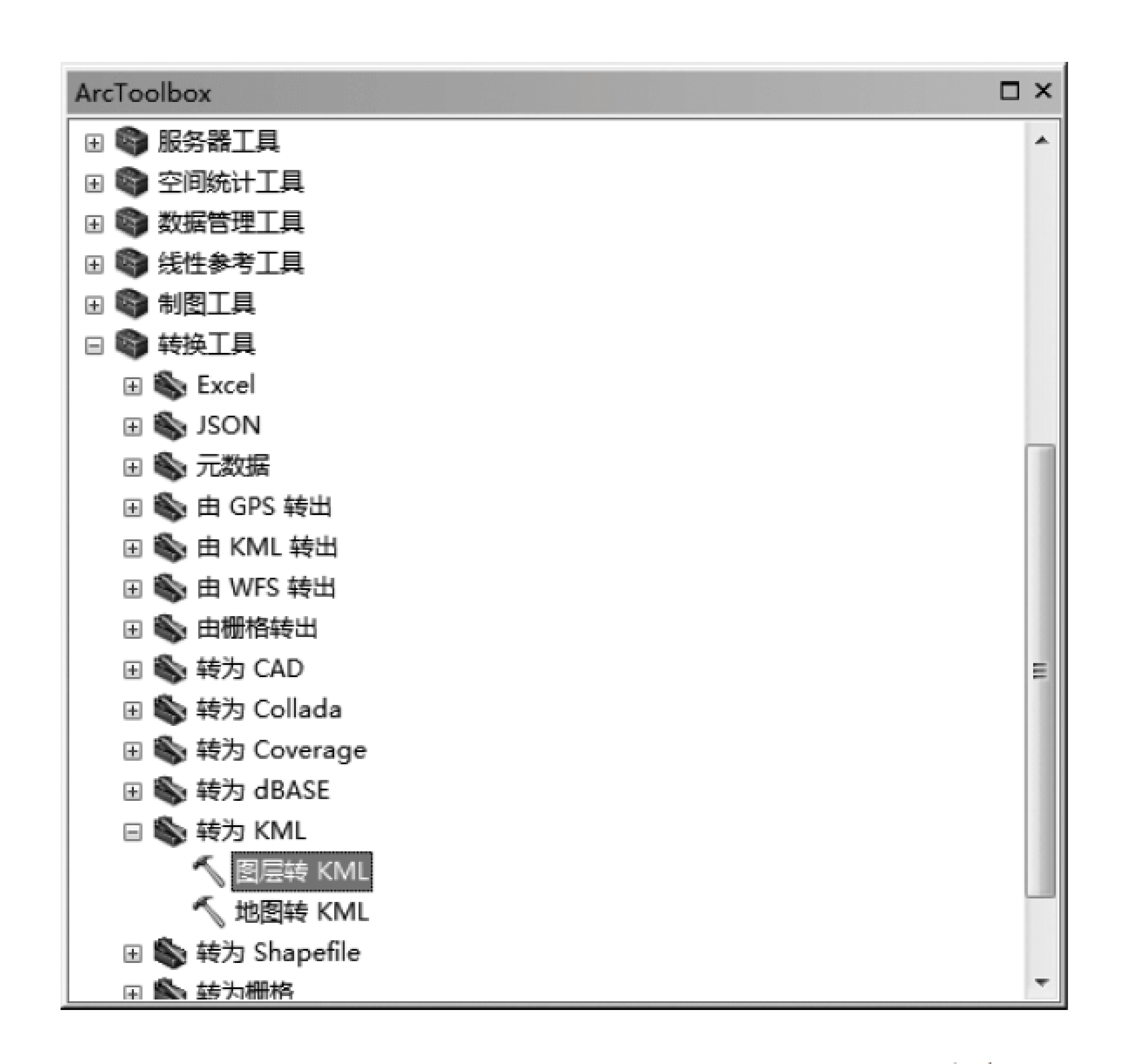




图 4-12　图层转换

(2) 逐项目逐地块核实

① 打开转换后的土地整治项目图层文件,如图 4-13 所示。双击某文件即可用 Google Earth 打开。

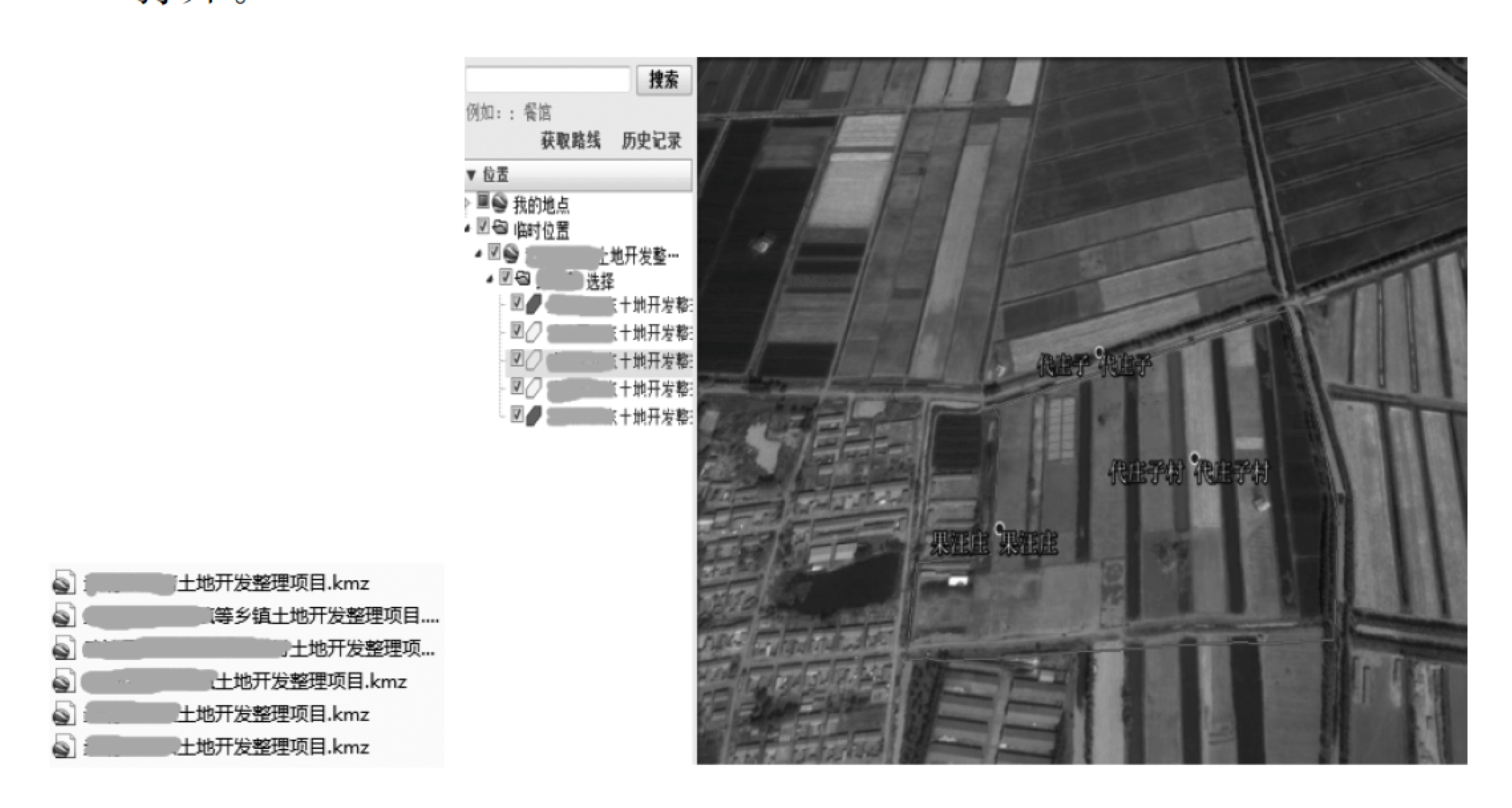


图 4-13　打开图层文件

② 双击图层或项目名称可获得该地块的具体信息,包括整治项目立项时间、验收时间和面积等,如图 4-14 所示。

(3) 通过时间轴工具,即可查看对比项目实施前后土地影像,如图 4-15 所示。

二、分析结果

审计组将 2008—2013 年被检查地区所有土地整治项目(不含高标准基本农田建设项目、农民自行开发耕地项目)逐一进行核实,共包括 94 个土地整治项目的 7 695 个地块,总面积 13 228 公顷。数据分析组按照下述原则判断是否为问题图斑:一是开发项目验收前后都是耕地;二是验收后还不是耕地;三是验收面积与实际开发面积不一致。分析结果显示,所有土地开发项目中疑似问题项目 87 个,包含 6 599 个地块,总面积 10 713 公顷,

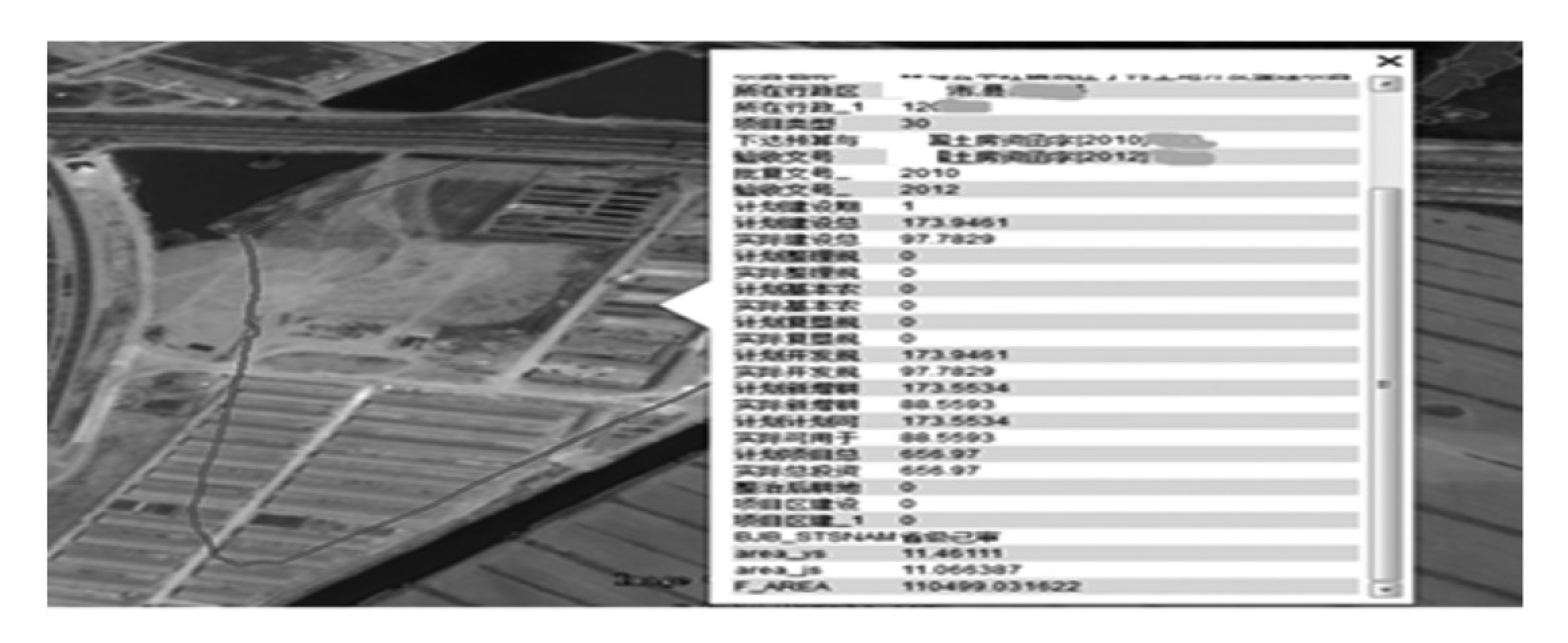


图 4-14　获得地块的具体信息



图 4-15　对比项目实施前后土地影像

占全部开发总面积的 81%。疑似问题图斑又可以分为以下五类。

1. 开发整理前后均为耕地，如图 4-16 所示。A 项目验收时间为 2008 年，面积 35.77 公顷。对比 2006 年卫片(左图)和 2014 年卫片(右图)，此块土地在开发整理前后均为耕地。




图 4-16　开发整理前后均为耕地

2．开发整理前为耕地，开发整理后为建设用地，如图 4-17 所示。B 项目验收时间为 2009 年，面积 5.82 公顷。对比 2008 年卫片（左图）和 2014 年卫片（右图），此块土地在开发整理前为耕地，在开发整理后变为建设用地。




图 4-17　开发整理后为建设用地

3．开发整理后撂荒，如图 4-18 所示。C 项目验收时间为 2009 年，面积 1.52 公顷。对比 2010 年卫片（左图）和 2014 年卫片（右图），此块土地在开发整理后又撂为荒地。

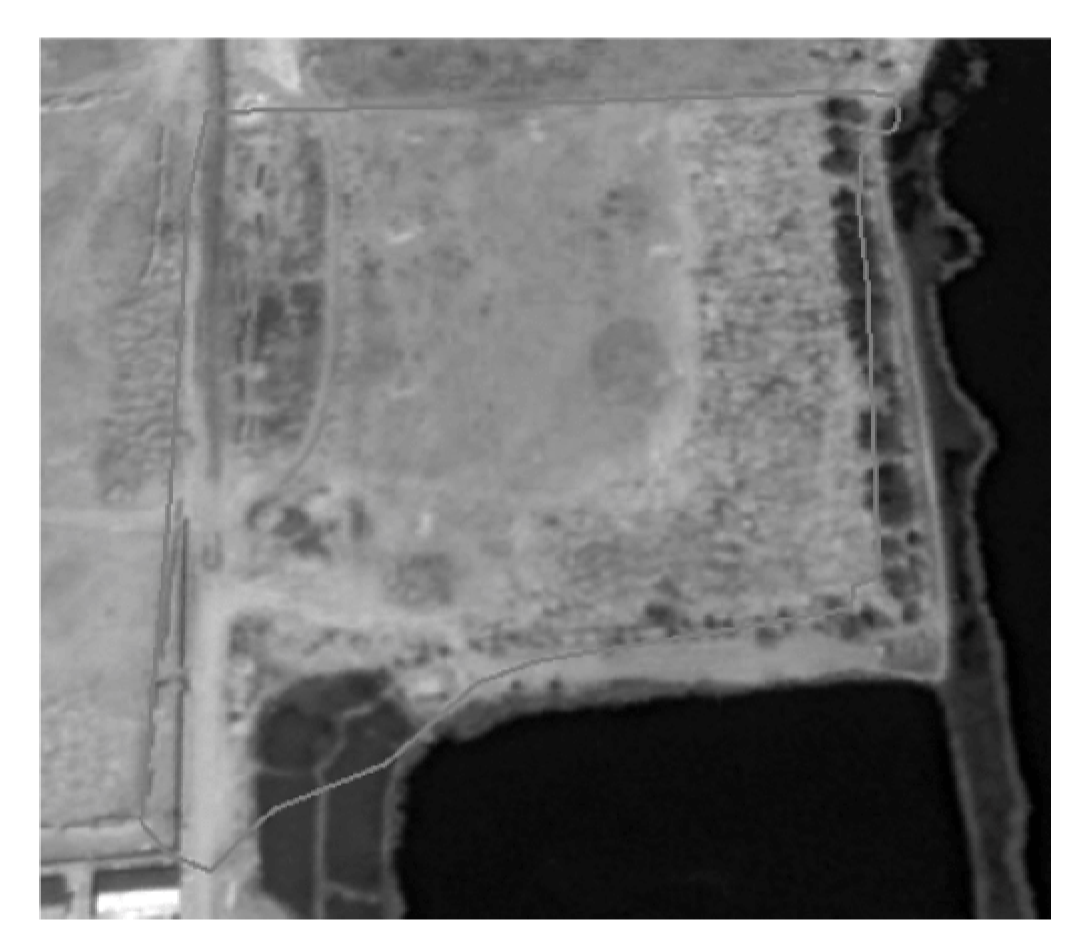



图 4-18　开发整理后撂荒

4．开发整理前为荒地，开发整理后亦为荒地，如图 4-19 所示。D 项目验收时间为 2010 年，面积 3.35 公顷。对比 2009 年卫片（左图）和 2014 年卫片（右图），此块土地在开发整理前为荒地，在开发整理后亦为荒地。

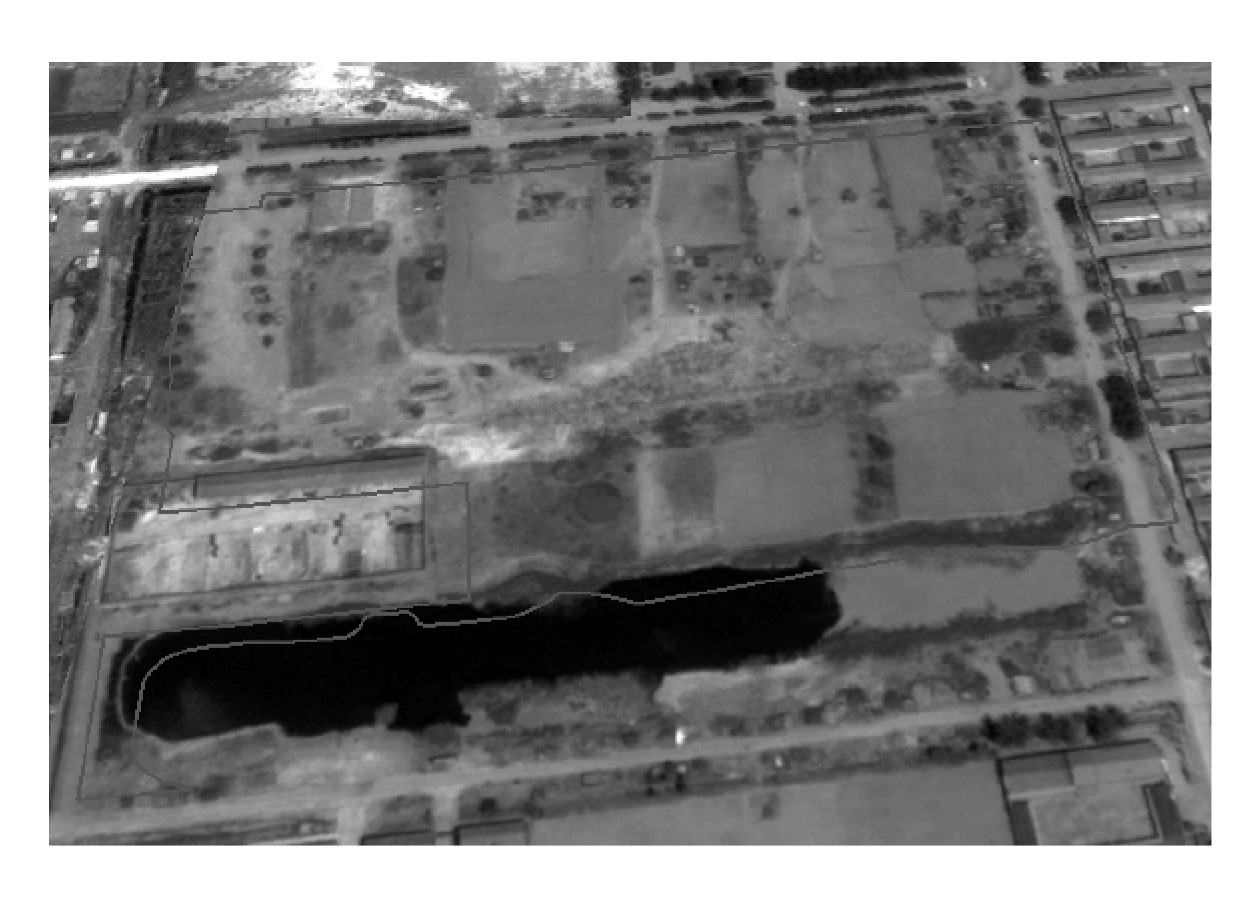



图 4-19　开发整理前后都为荒地

5. 开发整理项目重复申报,如图 4-20 所示。共有三个项目,偏左侧虚线和白线标识的两个区域为 2009 年开发整理项目,黑实线标识的区域为 2010 年开发整理项目,这个项目与 2009 年开发整理项目有重叠,表明项目重复申报,重复申报面积达 2.65 公顷。

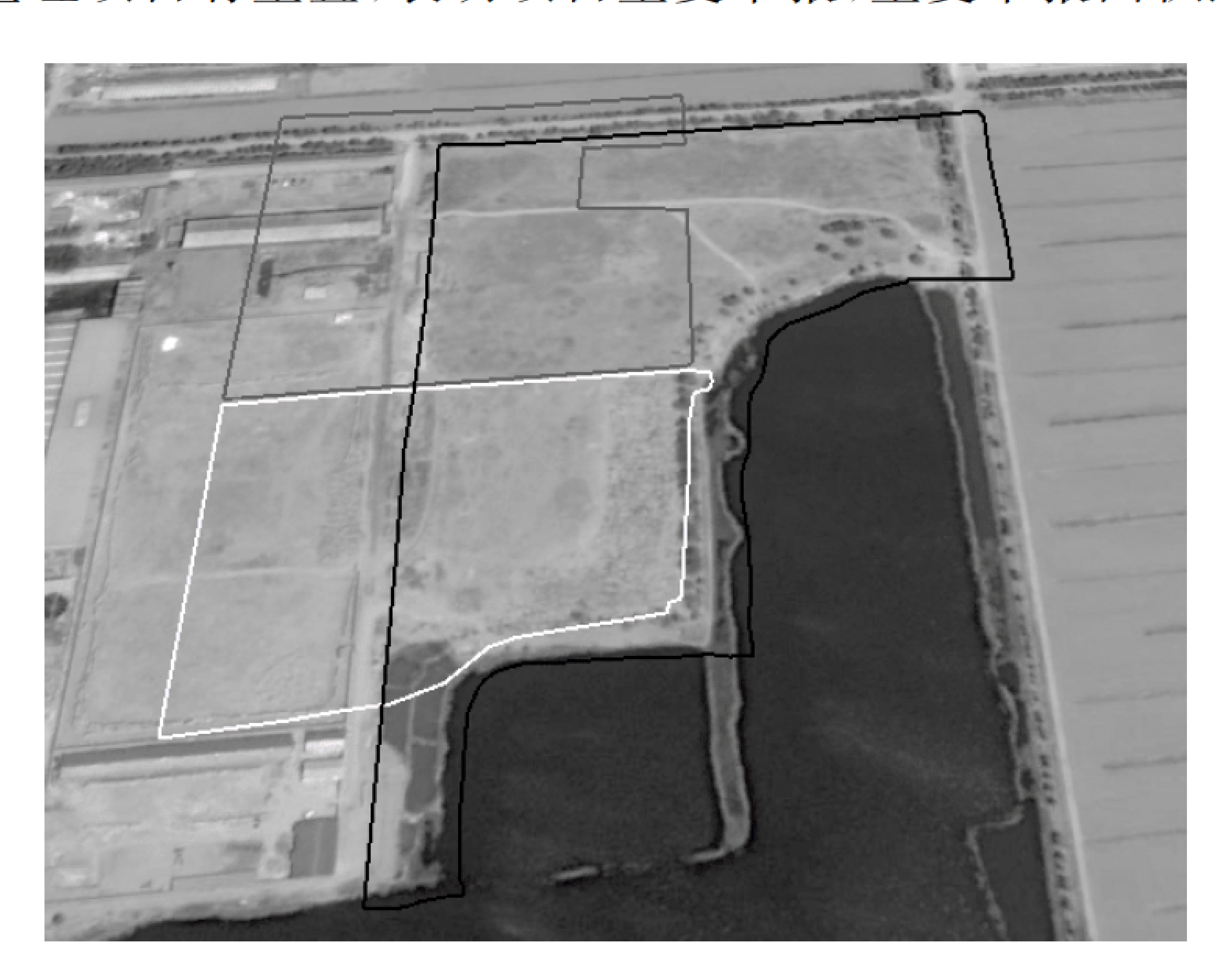


图 4-20　开发整理项目重复申报

审计组将发现的疑似问题图斑在地图中以红色标记,制成统一汇总的 Excel 表格,逐一核实,顺利完成了审计数据分析任务。

4.4　特征发现的一般过程

所谓过程,是指事物进行或事物发展所经过的程序。特征发现的一般过程是指为开展数据分析实施的步骤、程序,包括特征枚举、特征捕捉和特征分析三个步骤。特征发现的一般过程如图 4-21 所示。

需要强调的是特征发现是从分析性审计中间表开始的,是在分析性中间表的基础上进行数据分析、特征发现的过程。至于形成分析性中间表的方法和过程则不再赘述,读者可参考《审计分析模型算法(第 2 版)》(刘汝焯,北京,清华大学出版社)等书籍。

1. 特征枚举

特征枚举就是在特征发现过程中首先要尽量列举可能的特征表现。特征枚举需要一定的经验积累,就是要总结出什么样的线索会通过什么样的特征和方式表现出来。

但是有一种情况例外,即数据挖掘方法。海量数据到底会表现出什么特征,数据之间会有什么联系,在数据挖掘之前是不知道的,所以在利用数据挖掘的特征发现方法时,我们事前是无法进行特征枚举的。除此以外,其他特征发现方法都应该首先进行特征枚举。

2. 特征捕捉

传统的财政财务收支及经营管理等经济活动是以书面形式记载和反映的,各种违法违纪问题都会以书面的形式留下痕迹。在信息化条件下,电子数据成为财政财务收支和其他经济活动的主要记载和反映形式,违法违纪问题的痕迹隐藏在电子数据中。如何捕

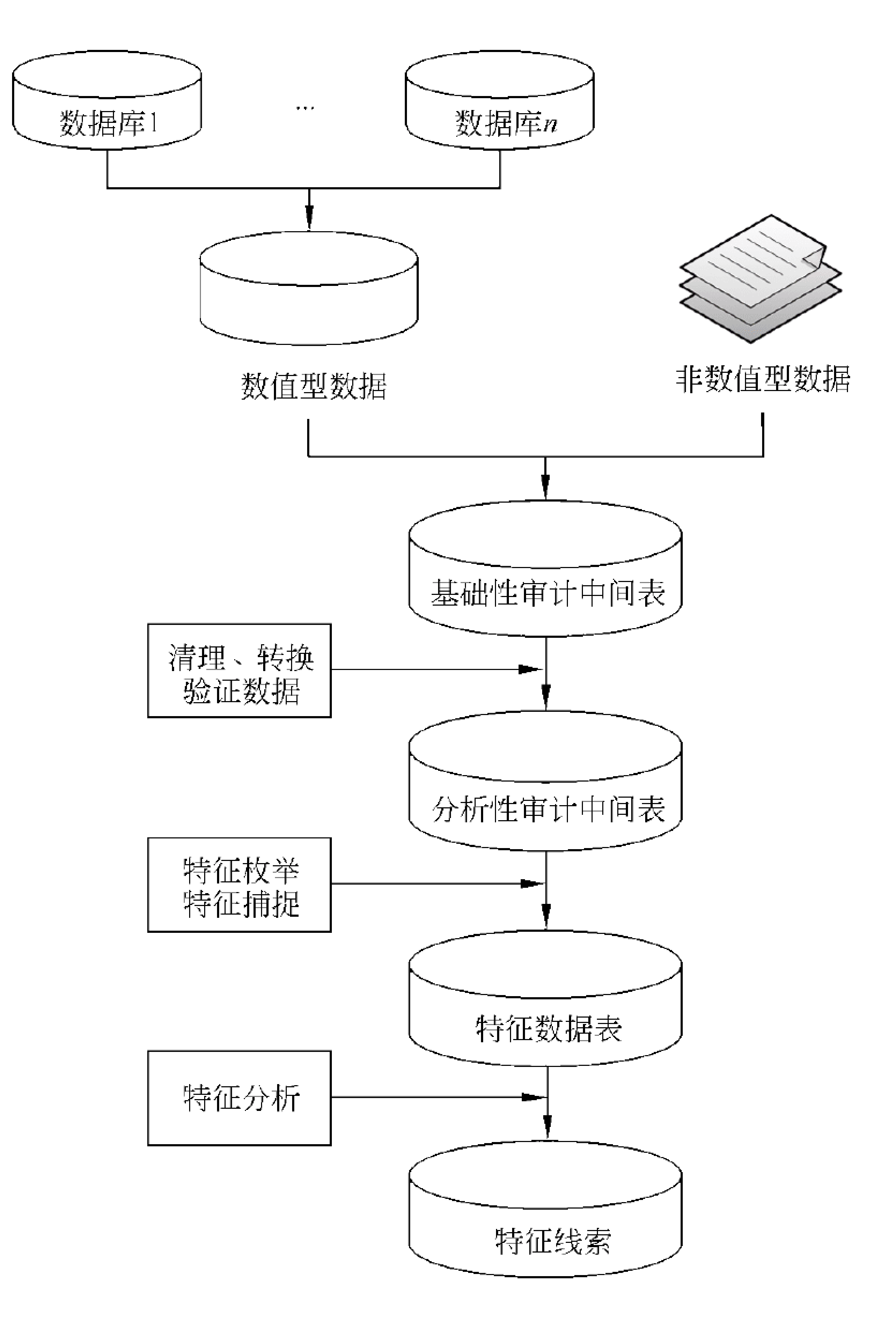


图 4-21 特征发现的一般过程

捉到这些特征？

特征捕捉就是通过运用计算机的查询功能或多维分析技术等相应方法来寻找符合相关特征的数据，或验证数据的发展趋势是否与通常的规律相一致的过程。数据特征隐藏在海量的数据中，特征捕捉是寻找、捕捉并显性化数据特征的过程。

3. 特征分析

特征分析就是根据线索分析取证。通过特征枚举和特征捕捉找出了符合特征表现的数据，对这些数据还需要进一步分析解读其中的意义。

实际上，特征捕捉和特征分析是两个互有渗透，不能截然分开的过程，特征捕捉前需要进行分析，如枚举的特征应该从哪些数据入手才能进行有效的捕捉、应该用什么方法才能有效地发现特征等。特征捕捉后更需要进行进一步的分析，因为有异常、符合特征表现的数据是否有问题，与之相关的经济活动是否违纪违规，还得依赖分析人员结合法律法规进行分析判断，需要引入分析人员的知识与经验。视频、音频数据的特征发现又有适合自身数据特征的新的特点。随着人工智能技术的深入发展，智能视频监控分析软件的功能也越来越强大。例如，用摄像头实时录像，用鼠标点击一个图像中的某一个人，这个人的

特征，如脸型、眼睛、身高、衣服的特征就被及时抓取，进入特征库，然后这个人就被实时跟踪。这种动态的特征发现，打开了一个更为广阔的应用空间。

参考文献

刘汝焯，等. 审计线索的特征发现[M]. 北京：清华大学出版社，2009.

第 5 章　大数据的可视化分析

本节通过两个案例演示用 Tableau 对数据进行可视化分析的过程，第一个案例为与审计业务相关的不良贷款分析，第二个案例为保险公司客户索赔与赔付分析。通过这两个案例演示可视化数据分析在不同领域的应用。

本章案例均使用 Tableau 10.0 版本进行分析演示。

5.1　不良贷款分析

贷款业务是目前各商业银行最基本的一项经营业务，数量多，涉及内容广，是商业银行资金运用的重要组成部分。大多数商业银行的收入主要来源于其发放的贷款，利润则在很大程度上取决于信贷业务量，信贷资产质量的好坏和业务量增长的快慢直接关系到银行的生存与发展。

银行在发放贷款时需要考虑贷款人的偿还能力，尽可能减少不良贷款的风险。本案例以法人贷款为例，分析找出不良贷款多或不良贷款率高的行业、地区及其经济类型等，为以后对贷款发放的监控提供科学的依据，降低不良贷款数量。

本案例使用的数据源为 SQL Server 中的“贷款数据库”，该数据库包含的数据表及结构如下：

- 主表_法人借款凭证表（机构编码，市行名称，支行名称，支行管辖机构名称，客户代码，借款凭证编号，贷款类别大类，贷款类别明细分类，贷款性质分类，币种，借款金额，借款日期，本凭证贷款余额，担保方式大类，担保方式明细，贷款四级分类大类，贷款四级分类明细，贷款五级分类，增量标志）
- 附表_法人基本信息表（客户代码，客户名称，法人代码，行业分类 1，行业分类 2，经济类型，经营状况）
- 代码表_经济类型代码表（经济类型大类，经济类型大类名称，经济类型明细，经济类型明细名称）
- 代码表_行业分类代码表（行业代码，行业名称）

5.1.1　数据准备

1. 连接数据源

在 Tableau 中建立数据源窗口，选择连接到“Microsoft SQL Server”，在选择数据库部分选择“贷款数据库”。此时连接数据源窗口如图 5-1 所示。

（1）选择数据表及表连接方式

在“表”列表框中首先分别双击“主表_法人借款凭证表”“附表_法人基本信息表”，然

后双击“代码表_经济类型代码表”，在弹出的“连接”窗口中，在左边的列表框中选择“附表_法人基本信息表”中的“经济类型”，在右边的“代码表_经济类型代码表”列表框中选择“经济类型明细名称”，设置好后的情形如图 5-2 所示。

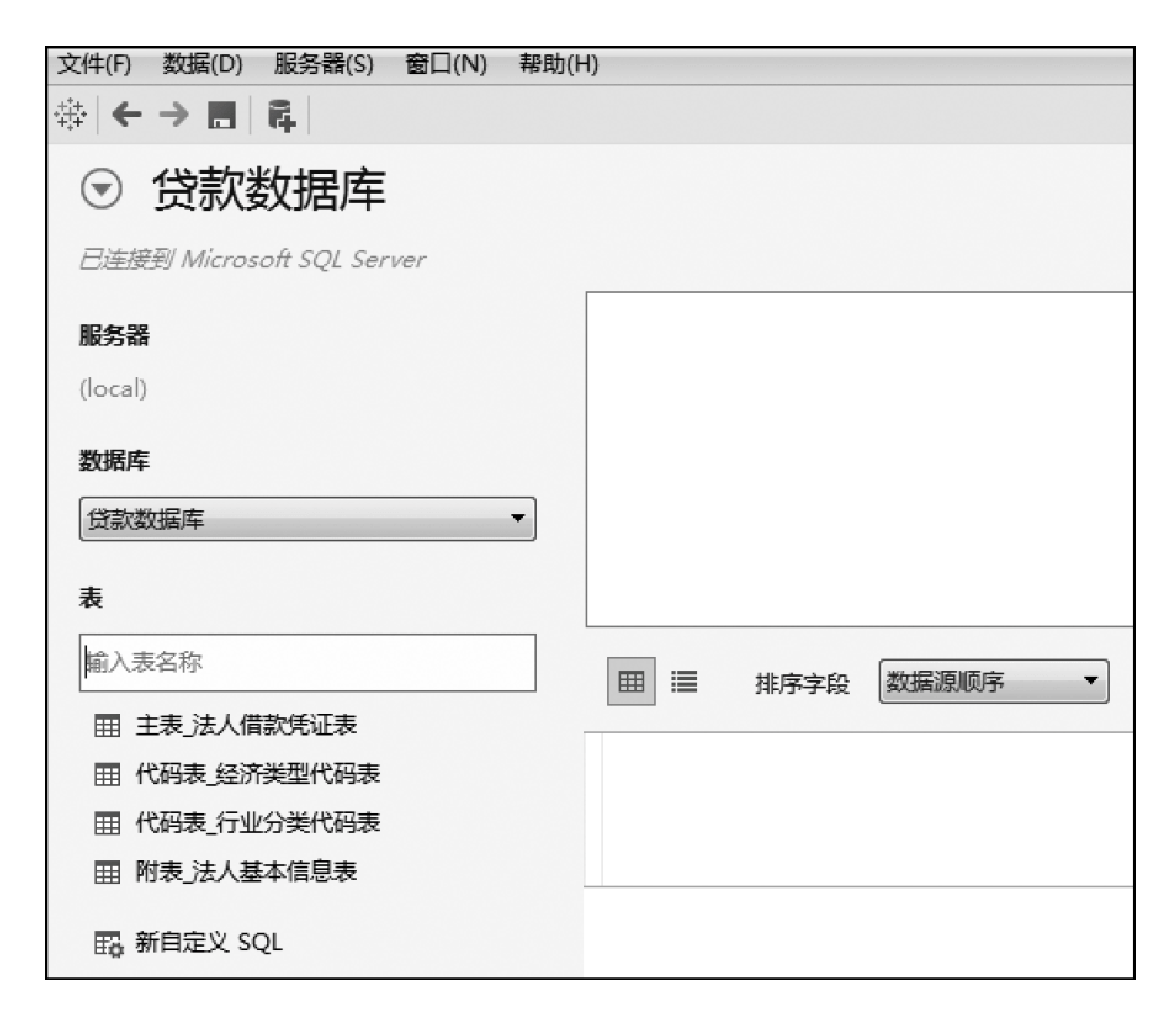


图 5-1　选择好连接的数据库后的窗口样式

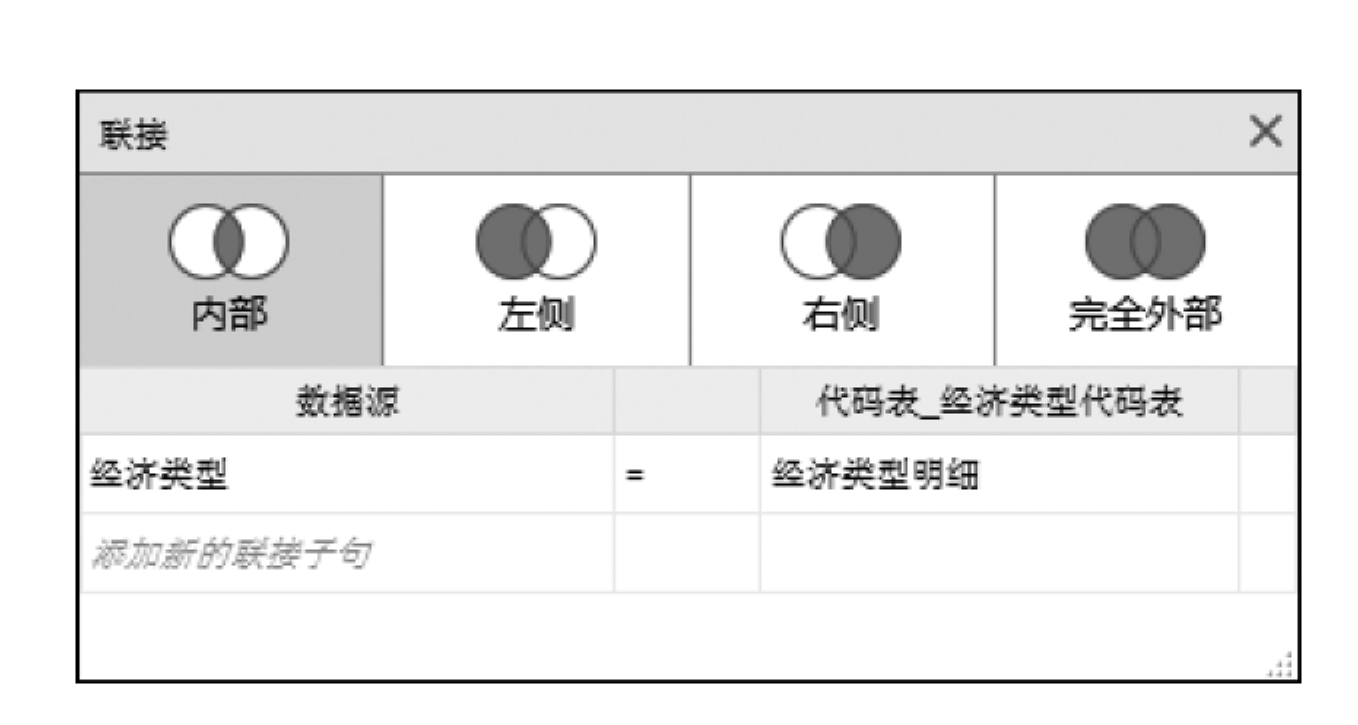


图 5-2　设置法人基本信息表与经济类型代码表之间的连接字段

最后双击“表”列表框部分的“代码表_行业分类代码表”，在弹出的“连接”窗口中，在左边的列表框中选择“附表_法人基本信息表”中的“行业分类 2”，在右边的“代码表_经济类型代码表”列表框中选择“行业代码”，设置好后的情形如图 5-3 所示。



图 5-3　设置法人基本信息表与行业分类代码表之间的连接字段

设置好数据源后各表的连接形式如图 5-4 所示。



图 5-4　设置好表连接条件后的连接样式

(2) 筛选数据

由于我们只分析币种为“人民币”且增量标志不为 4 和 6 的贷款数据，因此在进行数据分析之前，首先对数据源中的数据进行筛选。筛选方法如下：

单击建立数据源窗口右上角的“筛选”部分的 添加... ，弹出如图 5-5 所示的“编辑数据源筛选器”窗口。在此窗口中单击“添加”按钮，进入如图 5-6 所示的“添加筛选器”窗口。



图 5-5　“编辑数据源筛选器”窗口



图 5-6　选中“币种”字段

在“添加筛选器”窗口的“选择字段”列表框中选中“币种”，单击“确定”按钮，进入如图 5-7 所示的“筛选器[币种]”窗口，在此窗口中勾选“人民币”选项。单击“确定”按钮，关

闭“筛选器”窗口。



图 5-7　勾选“人民币”选项

按此方法设置“增量标志”的筛选条件：增量标志<>4 and 增量标志 <>6

设置好筛选条件后的“编辑数据源筛选器”形式如图 5-8 所示。单击“确定”按钮关闭该窗口，完成数据源的筛选。



图 5-8　设置好数据筛选条件后的筛选器窗口

2. 类型转换

由于源数据中的“借款日期”为字符串类型，为便于按日期进行数据分析，将“借款日期”类型改为“日期”。

3. 重命名字段

为方便分析和理解数据，将“本凭证贷款余额”重命名为“贷款总额”。

4. 创建计算字段

由于主要是分析不良贷款和不良贷款率，而这两类数据在“主表_法人借款凭证表”中是通过贷款类别来标识的，为分析方便，创建两个计算字段：不良贷款、不良贷款率。

(1) 不良贷款=

```
case [贷款五级分类]
   when '次级' then [贷款总额]
   else 0
end
+
case [贷款五级分类]
   when '可疑' then [贷款总额]
   else 0
end
+
case [贷款五级分类]
   when '损失' then [贷款总额]
    else 0
end
```

(2) 不良贷款率=SUM(不良贷款)/SUM(贷款总额)

5. 构建层次结构

(1) 担保方式：担保方式大类→担保方式明细

(2) 经济类型：经济类型大类名称→经济类型明细名称

(3) 行业分类：行业分类 1→行业名称

(4) 贷款类别：贷款类别大类→贷款类别明细分类

(5) 贷款四级分类：贷款四级分类大类→贷款四级分类明细

(6) 银行：市行名称→支行名称→支行管辖机构名称

5.1.2 各银行的不良贷款情况分析

1. 把握总体：各银行的贷款总额及不良贷款情况

分析目标：把握各银行的总体贷款情况及不良贷款情况，以及两者的宏观对比。

分析实现过程：

(1) 将“市行名称”拖放到“列”功能区，将“贷款总额”和“不良贷款”分别拖放到“行”功能区。

(2) 单击“行”功能区中“不良贷款”的下三角按钮，在弹出的菜单中选择“双轴”。在右侧纵坐标上右击鼠标，在弹出的菜单中选择“同步轴”。

(3) 在“标记”卡中，将“贷款总额”和“不良贷款”的图形均改为“区域”。

生成的分析图如图 5-9 所示。从图中可以很清晰地了解各银行的贷款总额与不良贷款的对比。

(4) 将维度中的“借款日期”拖放到“筛选器”，在弹出的“筛选器字段”窗口中，选中“年”(如图 5-10 所示)，我们这里按年分析不良贷款情况。单击“下一步”按钮，弹出如

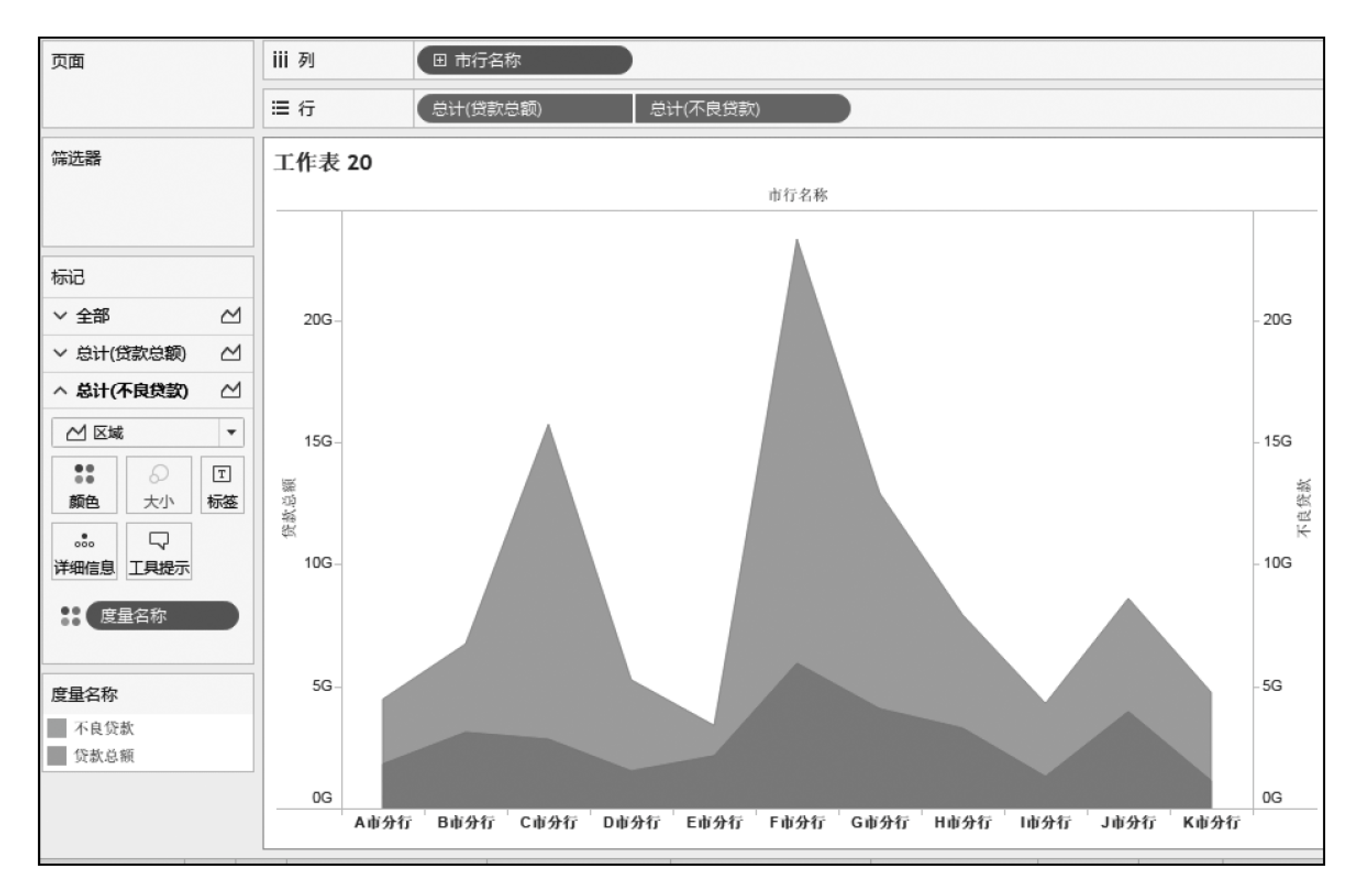


图 5-9　各银行贷款总额与不良贷款对比

图 5-11 所示的“筛选器”窗口，在此窗口中勾选“2010”前的复选框(假设分析 2010 年的贷款情况)。单击“确定”关闭该窗口。



图 5-10　在筛选器中选中“年”

(5) 单击“筛选器”窗格中“借款日期”的下三角按钮，在弹出的菜单中选择“显示筛选器”，将对借款日期的筛选显示在分析窗口中，便于指定要分析的年份。

最终的分析窗口如图 5-12 所示，可以勾选多个年份来分析若干年中各银行贷款总额与不良贷款情况。



图 5-11　勾选要分析的年份(2010)

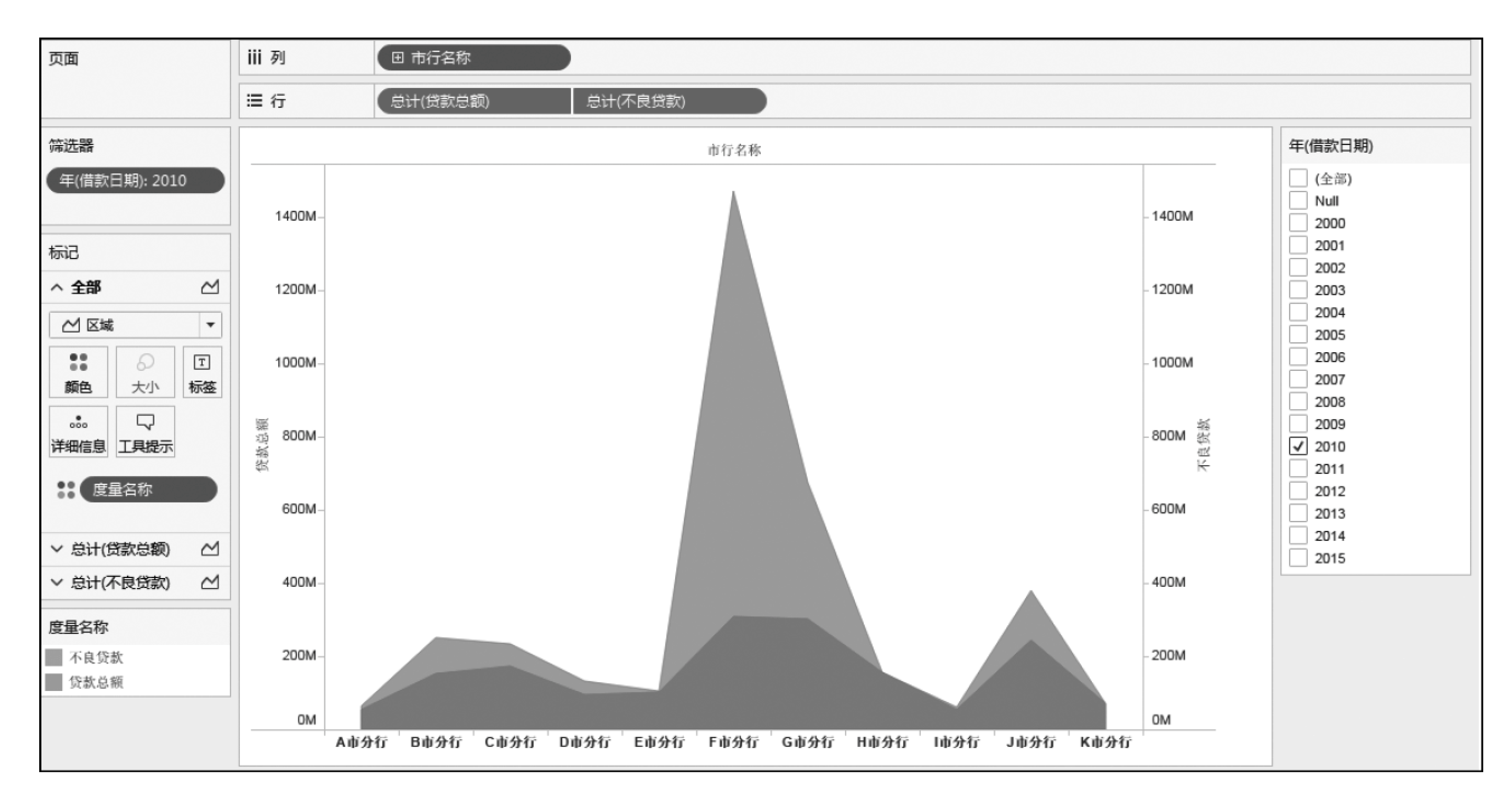


图 5-12　在指定年份中各银行贷款总额与不良贷款对比

2. 了解趋势：各银行不良贷款的历史变化情况

分析目标：分析若干年中各银行的不良贷款变化情况，以了解不良贷款的总体趋势是上升还是下降。

分析实现过程：

新建一个工作表,将“借款日期”拖放到“列”功能区,并将其粒度设置为“年”,将“不良贷款”拖放到“行”功能区,将“分行名称”拖放到“标记”卡中的“颜色”上,生成的分析图如图 5-13 所示。

从图中可以看到,各银行的不良贷款在 2008 年和 2009 年达到高峰,以后逐年减少。在 2008 年和 2009 年的不良贷款中以 F 市分行和 G 市分行尤为突出,不良贷款非常多。后面可以进一步分析 2008 年和 2009 年 F 市分行与 G 市分行中到底哪些支行和营业所的不良贷款发放比较多。

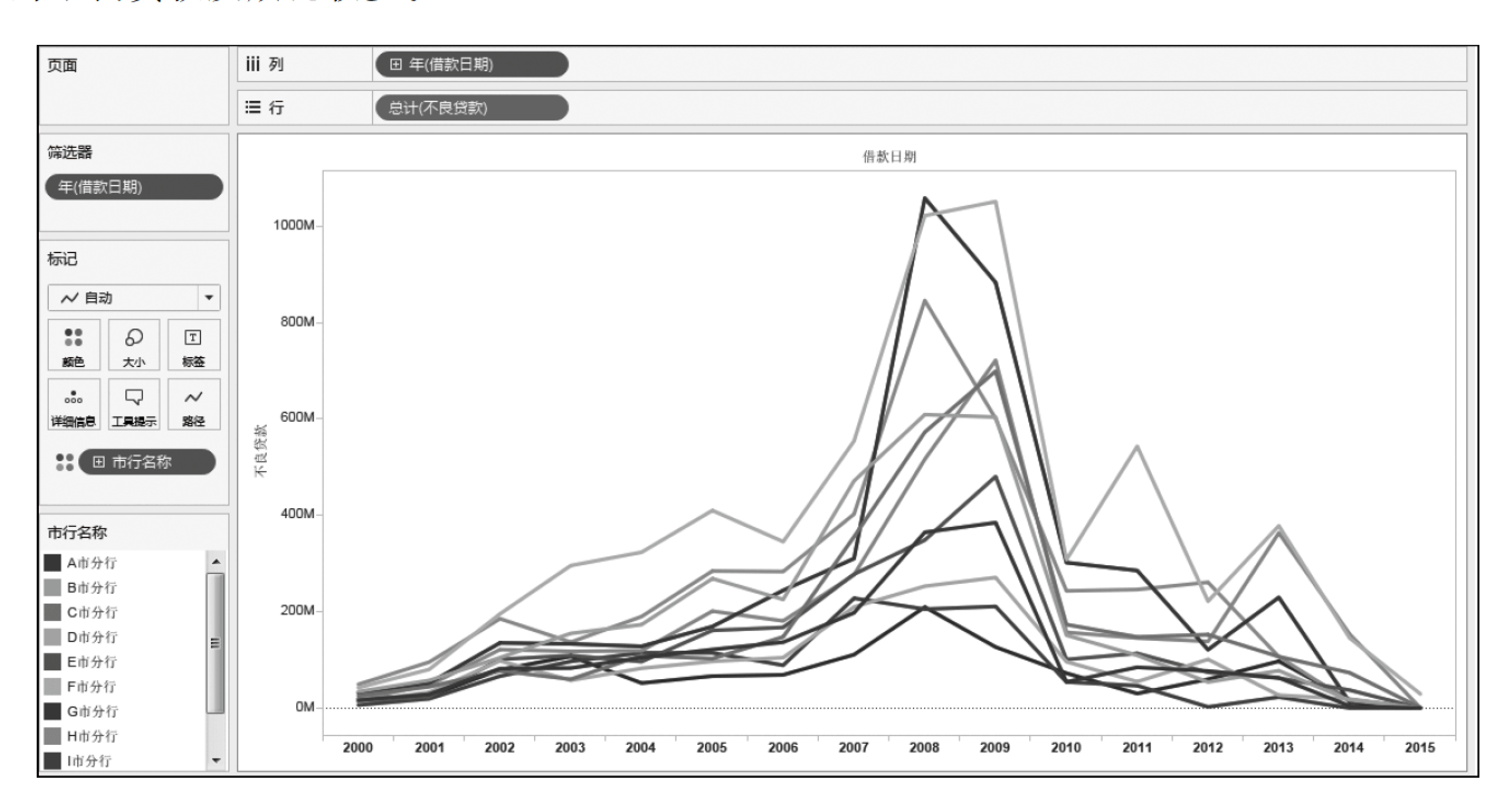


图 5-13　各银行不良贷款的历史变化情况

3. 了解趋势:各银行不同季度的不良贷款变化情况

分析目标:分析指定的若干年份中,每个市行每个季度的不良贷款变化情况。

分析实现过程:

新建一个工作表。将“借款日期”拖放到“行”功能区,设置其粒度为“年”。将“借款年份”拖放到“筛选器”,并选中“显示筛选器”,显示出年份筛选器,选中“2008”“2009”和“2010”。

将“市行名称”和“不良贷款”分别拖放到“列”功能区。

展开“列”功能区中的“年(借款日期)”到“季度”,并将“季度(借款日期)”拖放到“年(借款日期)”的前边。

最终产生的分析图如图 5-14 所示。从图中可以看出,2008—2010 年,各市行基本都是第 4 季度的不良贷款额呈下降趋势,而第 2 季度变化相对比较大。

4. 锁定重点:不良贷款率高的银行

分析目标:锁定不良贷款多及不良贷款率高的银行,以便对这些银行进行重点考察。

分析实现过程:

(1) 设置分析视图

新建一个工作表。在“标记”卡中,图形选择“方形”,将“市行名称”拖放到“标签”,将

“不良贷款”拖放到“大小”和“标签”，将“不良贷款率”拖放到“颜色”和“标签”。

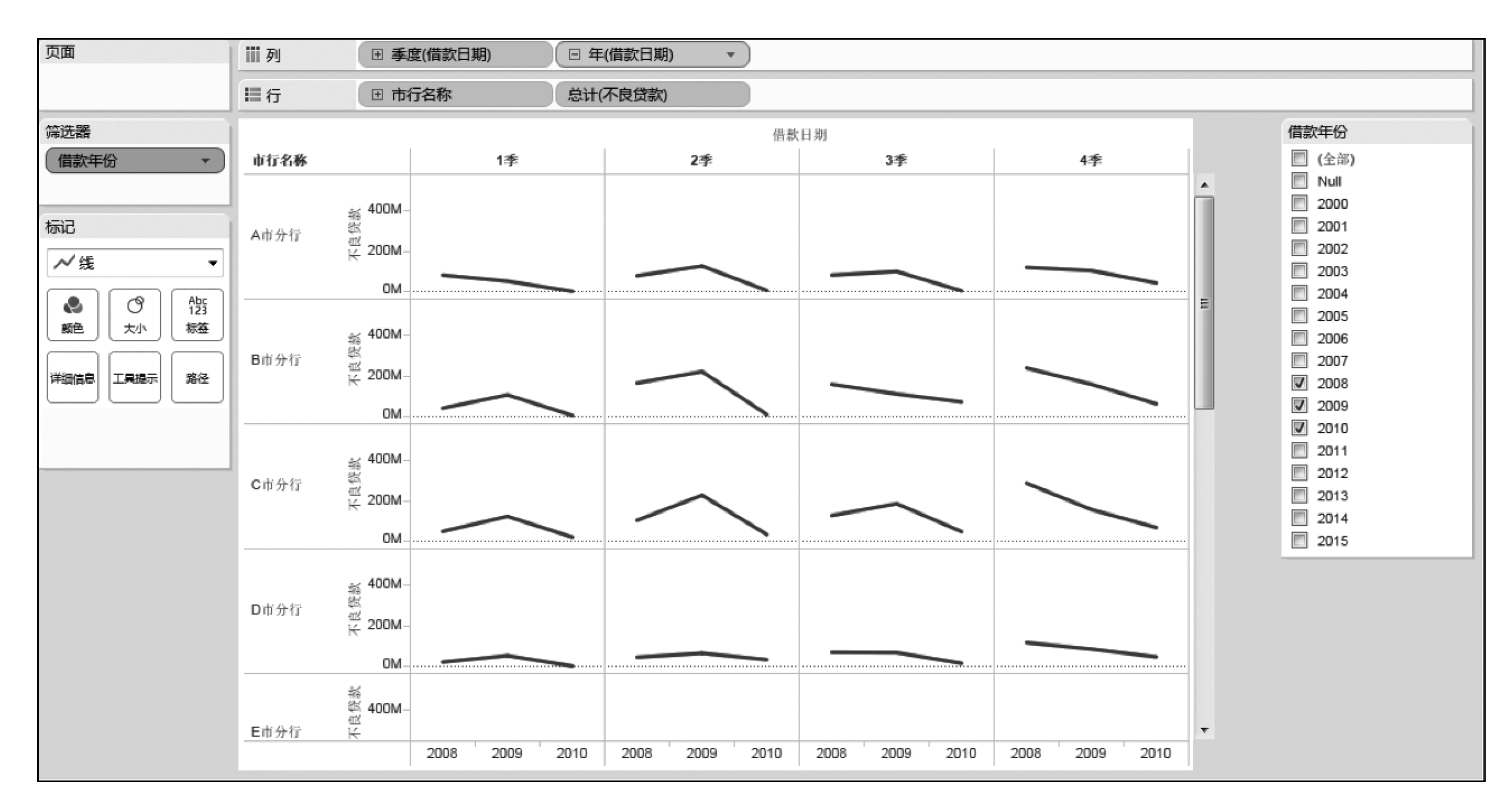


图 5-14　各银行不同季度的不良贷款变化趋势

（2）设置“不良贷款率”的显示格式，让其按百分比形式显示

单击“标签”上的“聚合(不良贷款率)”的下三角按钮，在弹出的菜单中选择“设置格式”，在出现的设置格式窗格(如图 5-15 所示)中，选中“区”选项卡，单击“默认值”部分的“数字”下拉列表框，弹出如图 5-16 所示的窗口，在该窗口中，在左边选择“百分比”，在右边的“小数位数”部分，设置小数位数为 2。



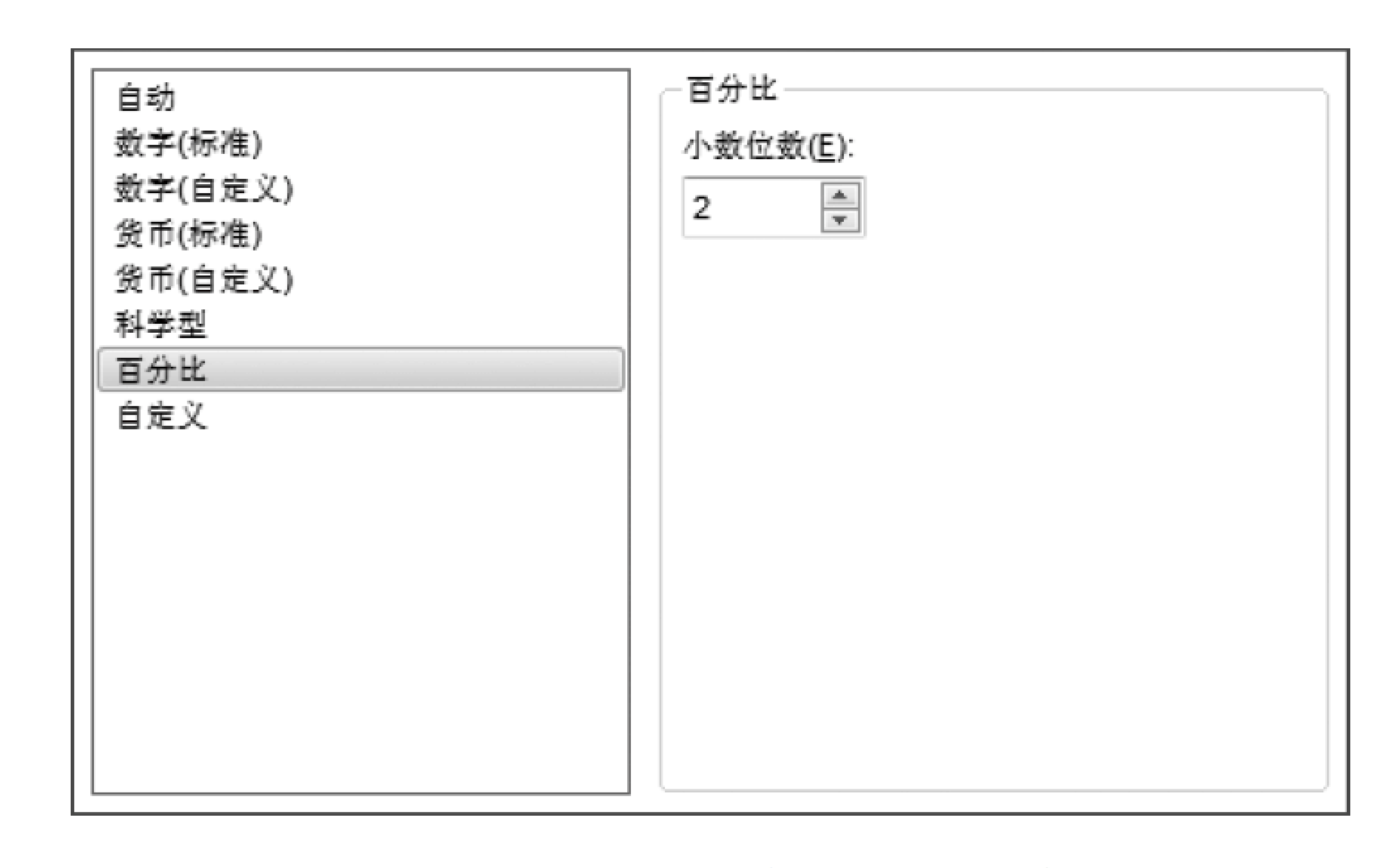


图 5-15　设置不良贷款率的显示格式

图 5-16　设置百分比显示形式

（3）设置筛选条件

将“借款日期”拖放到“筛选器”，设置选中条件为“年”，并选中“显示筛选器”，设置“借款日期”筛选条件的显示格式为“单选(下拉列表)”。

最终的分析结果如图 5-17 所示，该图显示了 2010 年各银行的不良贷款额及不良贷款率情况。图中方形的大小代表了不良贷款额的多少，方形越大，不良贷款额越高；颜色的深浅代表了不良贷款率的高低，颜色越深，不良贷款率越高。

从图 5-17 可以看到“F 市分行”虽然不良贷款额很高(307 957 327)，但其不良贷款率

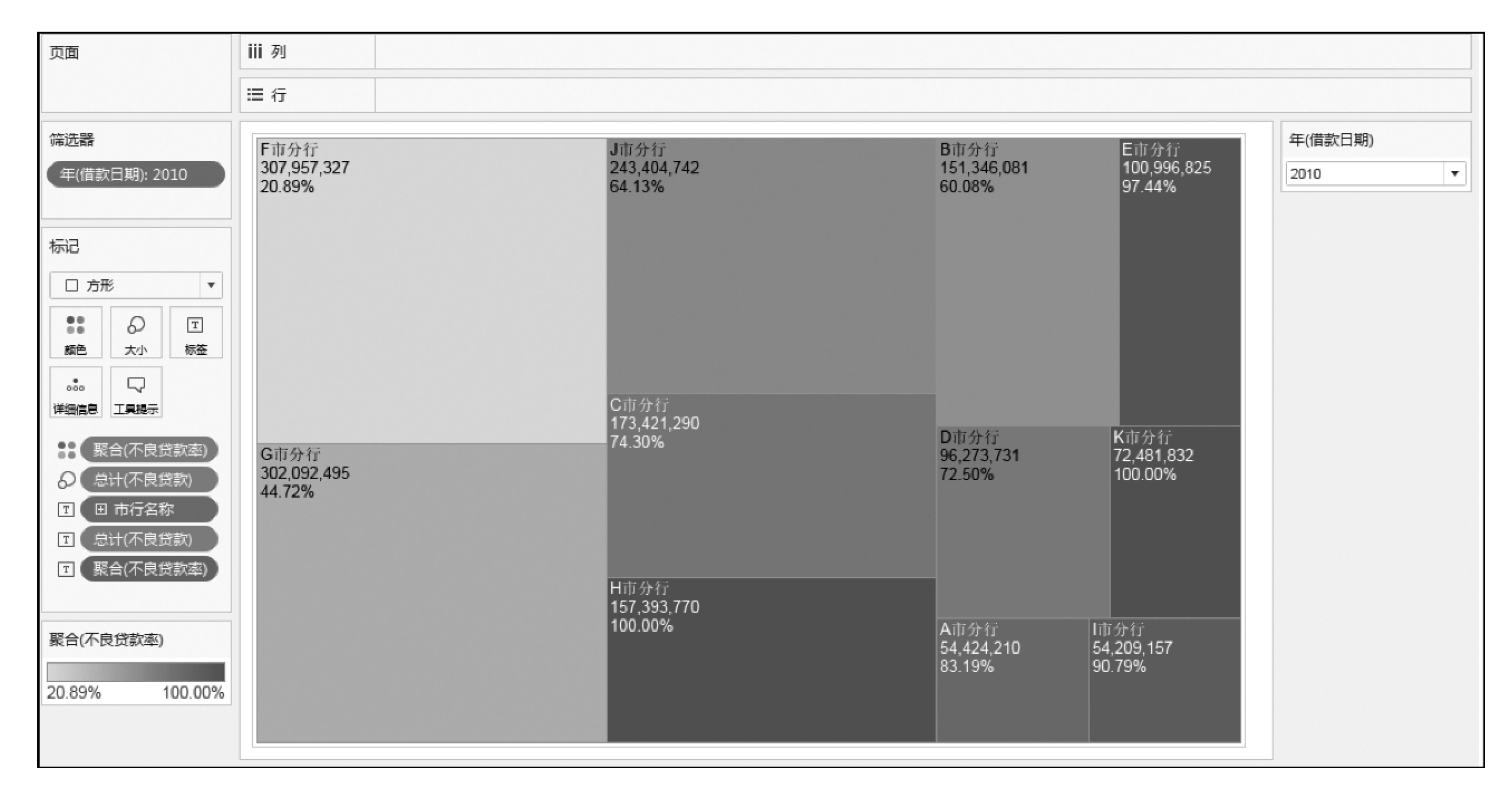


图 5-17　各银行不良贷款额及不良贷款率对比

并不高(20.89%)，而"K 市分行"的不良贷款额并不高(72 481 832)，但其不良贷款率却很高，达到了 100.00%。

通过在筛选器中指定不同的借款年份，可分别查看每年各银行的不良贷款额和不良贷款率。

5. 锁定重点：各银行的不良贷款占比

分析目标：分析各市行的不良贷款在当年总的不良贷款中所占的比例，找出不良贷款占比高的银行，便于后续对这些银行进行深入分析。

分析实现过程：

(1) 创建计算字段

- 借款年份＝year(借款日期)，并将"借款年份"转换为维度属性。
- 按年不良贷款额＝{ fixed [借款年份] ：sum([不良贷款])}
- 各行不良贷款占比＝sum([不良贷款])/sum([按年不良贷款额])

(2) 设置分析视图

新建一个工作表。在"标记"卡上，选择图形为"饼图"，将"市行名称"分别拖放到"颜色"和"标签"上，将"各行不良贷款占比"分别拖放到"角度"和"标签"上。

修改"各行不良贷款占比"的显示格式为"百分比"，小数点后 2 位。

(3) 设置筛选条件

将"借款年份"拖放到"筛选器"，并显示筛选器，勾选"2009"年(假设希望查看 2009 年各市行的不良贷款占比情况)。

分析结果如图 5-18 所示，从图中可以看到，"F 市分行"和"G 市分行"的不良贷款占比较多，分别占到了当年总的不良贷款的 17.42%和 14.64%。

(4) 更进一步，我们可以只筛选出不良贷款占比最多的前 N 家市行，比如占比最高的前 5 家市行，这可通过设置参数实现。

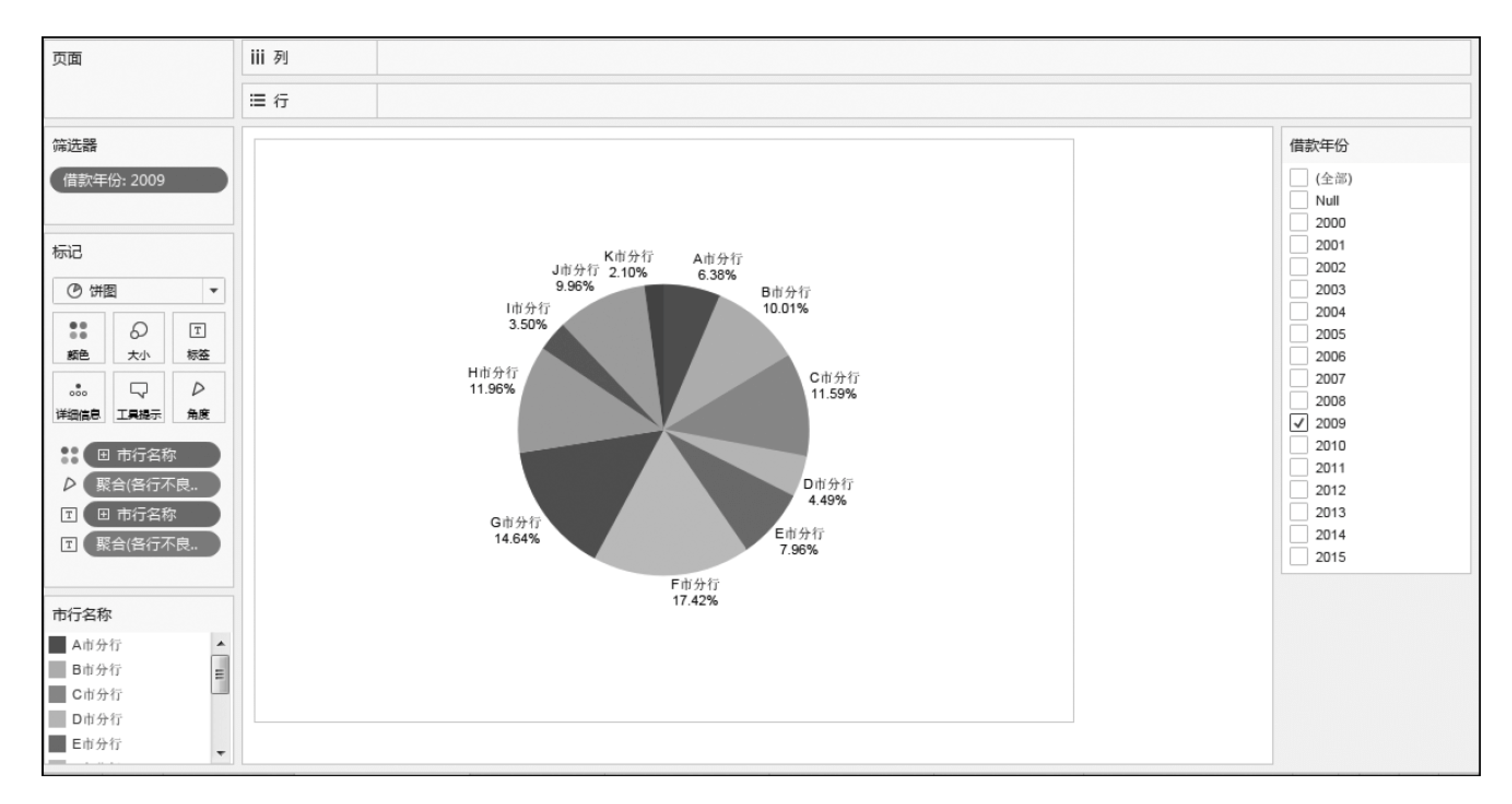


图 5-18　各市行的不良贷款占比分析

在“市行名称”上右击鼠标，在弹出的菜单中选择“创建”→“集”，弹出如图 5-19 所示的“创建集”窗口。在该窗口中，在“名称”文本框中输入集的名称“占比高的市行”，单击“顶部”选项卡。

图 5-19　“创建集”窗口

在“创建集”窗口的“顶部”选项卡中，选择“按公式”单选按钮，并在“依据”列表框中选择“创建新参数”。

在弹出的“创建参数”窗口（如图 5-20 所示）中，在“名称”文本框中输入参数的名称

创建参数
名称(N): 占比前N　注释(C) >>
属性
数据类型(T): 整数
当前值(V): 5
显示格式(F): 自动
允许的值(W): 全部(A)　列表(L)　范围(R)
值范围
最小值(U): 1　从参数设置(M)
最大值(X): 11　从字段设置(E)
步长(Z): 1
确定　取消

图 5-20　“创建参数”窗口

(这里是：占比前 N)，在“当期值”文本框中设置当期值为 5，表示默认显示不良贷款占比高的前 5 家市行。在“值范围”部分设置最小值为 1，最大值为 11(因为在贷款数据中只有 11 家不同的市行)。单击“确定”按钮，关闭“创建参数”窗口，接下来单击“创建集”窗口上的“确定”按钮，完成对集的创建。

将集“占比高的市行”拖放到“筛选器”中，单击参数“占比前 N”的下三角按钮，在弹出的菜单中选择“显示参数控件”。分析视图的形式如图 5-21 所示。

在如图 5-21 所示的分析图中，通过调整“占比前 N”参数控件可设置显示不良占比高的前若干家市行。通过“借款年份”筛选条件，可查看指定年份中不良贷款高的前若干家市行。

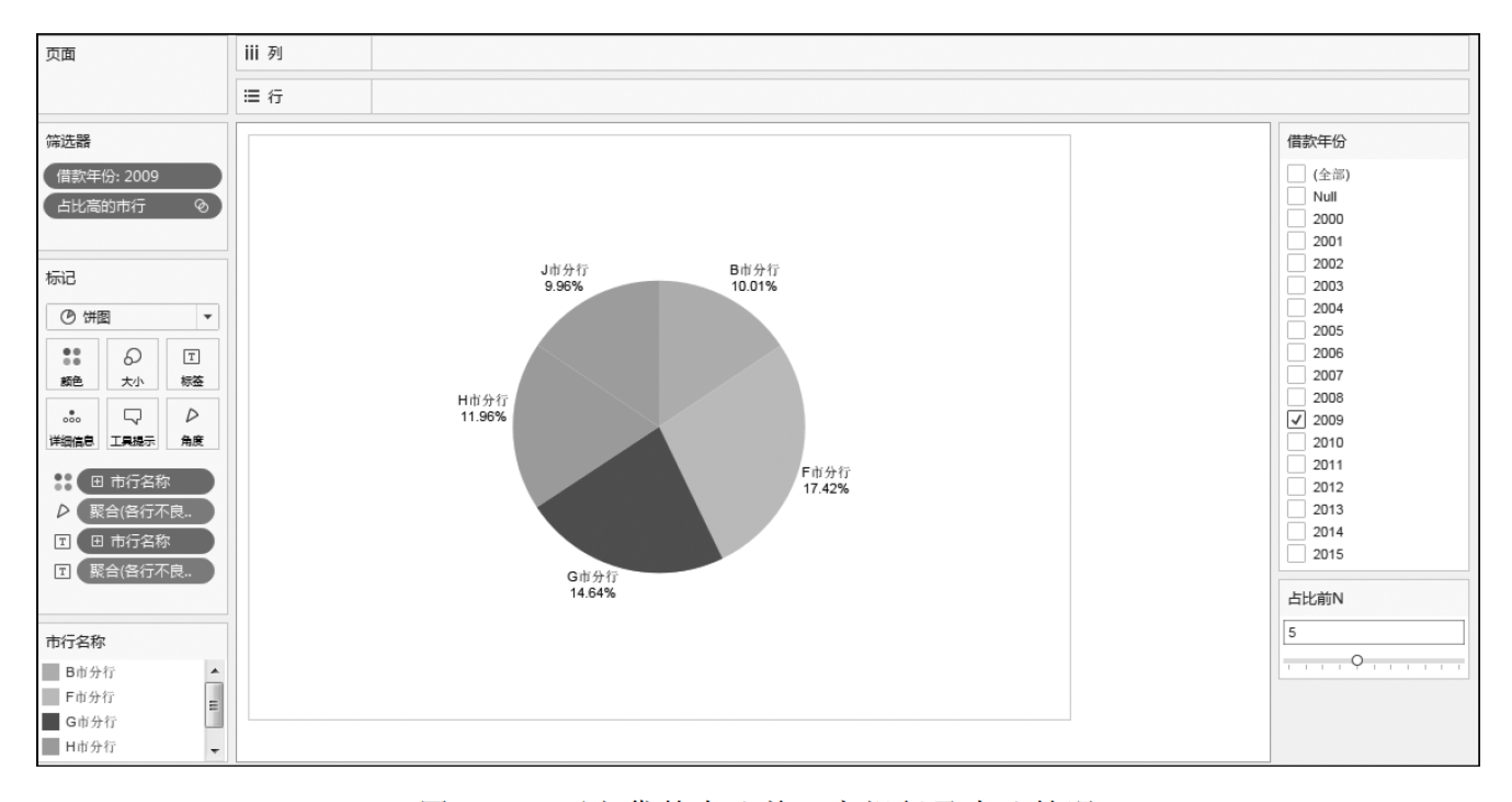


图 5-21　不良贷款占比前 5 家银行及占比情况

6．深入分析：发放不良贷款多的分支机构

分析目标：深入分析不良贷款多的两个市行管辖的所有分支机构的不良贷款发放情况。

从图 5-13 和图 5-17 可看到 2008 年和 2009 年不良贷款达到高峰，而这两年又以“F 市银行”和“G 市银行”的不良贷款最多。

分析实现过程：

(1) 设置分析视图

新建一个工作表。将“支行名称”拖放到“标记”卡的“颜色”和“标签”，将“不良贷款”拖放到“大小”和“标签”，将“支行管辖机构名称”拖放到“标签”。

在“智能显示”卡上选择“填充气泡图”。

(2) 设置筛选条件

将“市行名称”“借款日期”“不良贷款率”“不良贷款”分别拖放到“筛选器”中，并显示这些筛选器。

设置筛选条件：

- 市行名称＝F 市分行
- 借款日期＝2008 年
- 不良贷款率＞＝0.5
- 不良贷款总额＞＝40,000,000

生成的分析视图如图 5-22 所示。

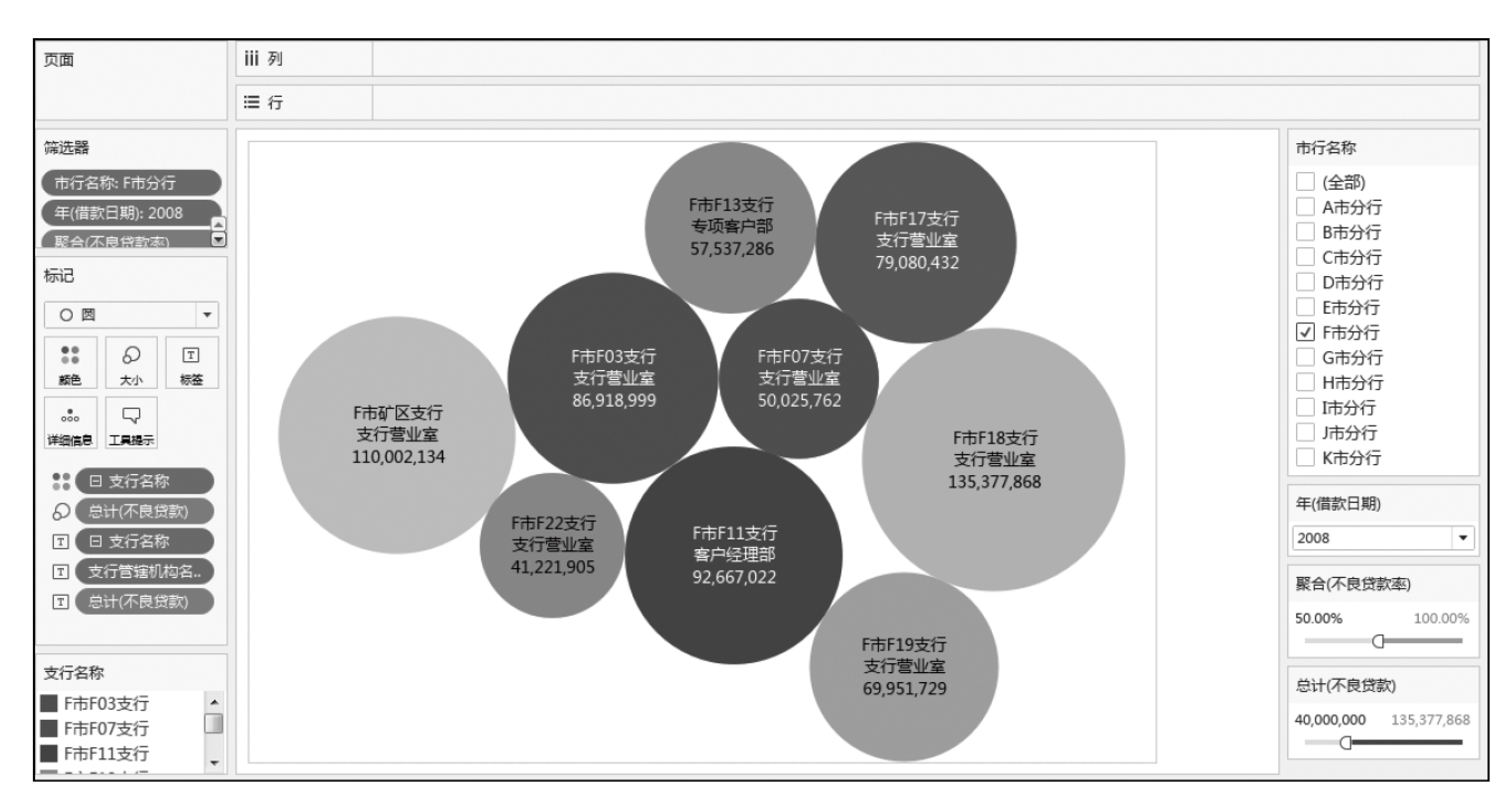


图 5-22　不良贷款额高的分支机构

在分析视图上右击鼠标，在弹出的菜单中选择“查看数据”，可以以表格形式查看视图中的数据，如图 5-23 所示。

单击图 5-23 中的“基础”选项卡，可对数据进行钻取，查看对应的更详细数据，如图 5-24 所示。单击图 5-24 上的“全部导出”按钮，可将数据导出到 Excel 格式的文件中。



查看数据

☑ 显示别名(S)　　复制(C)　全部导出(E)

支行名称	支行管辖机构名称	不良贷款
F市矿区支行	支行营业室	110,002,134.43
F市F22支行	支行营业室	41,221,905.24
F市F19支行	支行营业室	69,951,728.67
F市F18支行	支行营业室	135,377,867.68
F市F17支行	支行营业室	79,080,431.92
F市F13支行	专项客户部	57,537,286.28
F市F11支行	客户经理部	92,667,022.01
F市F07支行	支行营业室	50,025,761.73
F市F03支行	支行营业室	86,918,998.76

摘要　基础　　9 行

图 5-23　查看分析视图对应的数据

查看数据

744 行　☑ 显示别名(S)　☐ 显示所有字段(F)　复制(C)　全部导出(E)

借款日期	市行名称	支行名称	支行管辖机构名称	贷款五级分类	不良贷款	不良贷款率	贷款总额
2008/3/30	F市分行	F市矿区支行	支行营业室	可疑	200,000.00	1.00000	200,000.00
2008/4/20	F市分行	F市F03支行	支行营业室	损失	800,000.00	1.00000	800,000.00
2008/11/25	F市分行	F市F07支行	支行营业室	损失	1,763,000.00	1.00000	1,763,000.00
2008/11/25	F市分行	F市F07支行	支行营业室	损失	1,000,000.00	1.00000	1,000,000.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	33,500.00	1.00000	33,500.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	200,518.00	1.00000	200,518.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	10,540.00	1.00000	10,540.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	698,452.00	1.00000	698,452.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	40,000.00	1.00000	40,000.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	33,420.00	1.00000	33,420.00
2008/4/6	F市分行	F市F19支行	支行营业室	损失	849,392.00	1.00000	849,392.00

摘要　基础　　744 行

图 5-24　查看详细数据

5.1.3　各经济类型的企业的不良贷款情况分析

1. 锁定重点：不良贷款多的企业的经济类型

分析目标：分析不良贷款多的企业的经济类型，锁定需要深入分析的经济类型。

分析实现过程：

（1）设置分析视图

新建一个工作表。将“借款日期”和“不良贷款”拖放到“列”功能区，将“经济类型大类名称”拖放到“行”功能区和“标记”卡中的“颜色”上。将“不良贷款”拖放到“标记”卡的“标签”上。

（2）设置筛选条件

将“借款年份”计算字段拖放到“筛选器”，并选中“显示筛选器”，勾选“2008”、“2009”和“2010”。

将“不良贷款”拖放到“筛选器”，并选中“显示筛选器”，将“不良贷款”的值设置为：400,000,000。

至此,生成的分析图如图 5-25 所示。从图中可以看到 2008—2010 年不良贷款最多的企业分别是:集体企业、国有企业和私营企业。其中“集体企业”的不良贷款在这三年都是最多的;“国有企业”2008 年与“集体企业”的不良贷款差不多,但在 2009 年减少很多,在 2010 年已没有不良贷款;“私营企业”居第三,也是 2008 年和 2009 年有不良贷款,2010 年已没有了不良贷款。以后可以重点对“集体企业”进行分析。



图 5-25　不良贷款多的经济类型企业

2. 关联分析:给“集体企业”发放不良贷款的银行

分析目标:从图 5-25 的分析结果已经知道,“集体企业”的不良贷款额很高,是我们进一步分析的重点。下面分析给“集体企业”发放不良贷款的银行及不良贷款发放情况。

(1) 设置分析视图

新建一个工作表。将“借款年份”拖放到“列”功能区,并将粒度设置为“年”,将“不良贷款”拖放到“行”功能区,将“市行名称”拖放到“标记”卡的“颜色”上。

(2) 设置筛选条件

将“经济类型大类名称”拖放到“筛选器”,并只勾选“集体企业”。

将“借款年份”拖放到“筛选器”,取消对年份为 null 的勾选。

将“不良贷款”拖放到“筛选器”并选中“显示筛选器”,设置不良贷款的取值为 100,000,000,只分析不良贷款超过 100,000,000 的数据。

产生的分析视图如图 5-26 所示。从图中可以看到给“集体企业”发放不良贷款的情况。2008 年和 2009 年发放不良贷款最多,2008 年“G 市分行”发放的不良贷款最多,2009 年“F 市分行”发放的不良贷款最多。而且“F 市分行”2000—2009 年连续给“集体企业”发放了不良贷款,“G 市分行”是 2005—2010 年连续给“集体企业”发放了不良贷款。

以后即可重点考察“G 市银行”和“F 市银行”在 2008 年和 2009 年具体对哪些集体企业发放了较多的不良贷款。

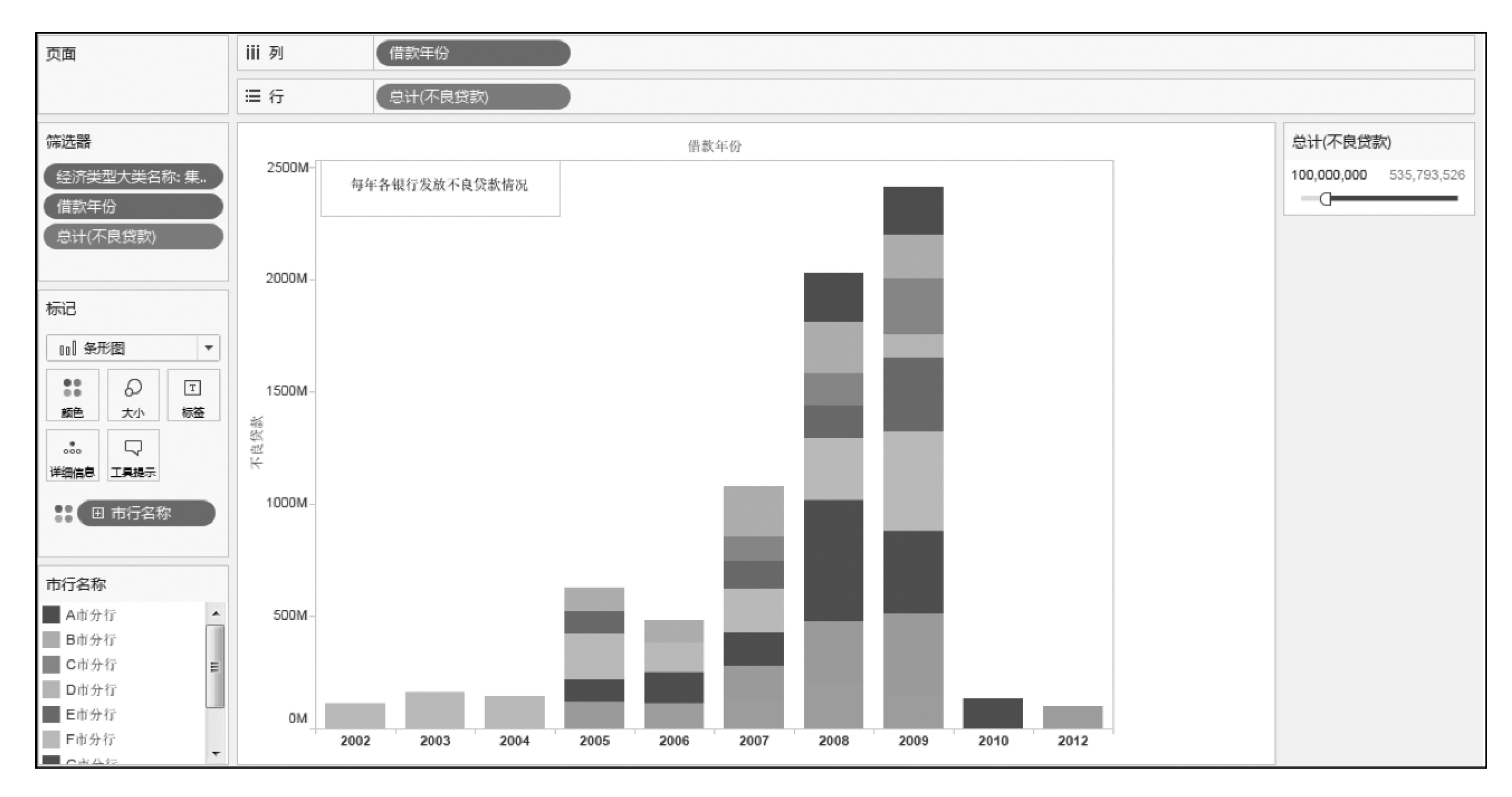


图 5-26　各银行历年给集体企业发放的不良贷款情况

3. 深入分析:“集体企业”中获得不良贷款多的企业

分析目标:在了解了 2008 年和 2009 年“G 市分行”和“F 市分行”对集体企业发放的不良贷款较多之后,接下来可深入分析这些不良贷款都发放给了哪些企业。

假设我们要找出 2008 年“G 市分行”“F 市分行”和“H 市分行”给集体性质的企业发放不良贷款最多的前 5 家企业的名称和所发放的不良贷款总额。

(1) 创建计算字段

排名=index()

单击“排名”字段上的下三角按钮,在弹出的菜单中选择“转换为离散”。

(2) 设置分析视图

新建一个工作表。将“市行名称”和“客户名称”分别拖放到“行”功能区,将“不良贷款”拖放到“列”功能区。

将“排名”拖放到“行”功能区中“市行名称”和“客户名称”之间,然后单击“排名”上的下三角按钮,在弹出的菜单中选择“编辑”,弹出如图 5-27 所示的定义计算字段窗口,在此窗口中单击“默认表计算”,弹出“表计算[排名]”窗口,在此窗口中单击“根据以下因素计算”下拉列表框,选择“高级”,弹出如图 5-28 所示的“高级”窗口。



图 5-27　定义计算字段窗口

在弹出的如图 5-28 所示的“高级”窗口中,分别将“分区”框中的“市行名称”和“客户

名称”添加到“寻址”框中，在“排序”部分，选中“字段”单选按钮，选择“不良贷款”和“总计”，在排序方式部分选择“降序”，以查看不良贷款合计最多的企业。设置好后的窗口形式如图 5-29 所示。单击“确定”按钮，回到“表计算”窗口。



图 5-28 “高级”窗口

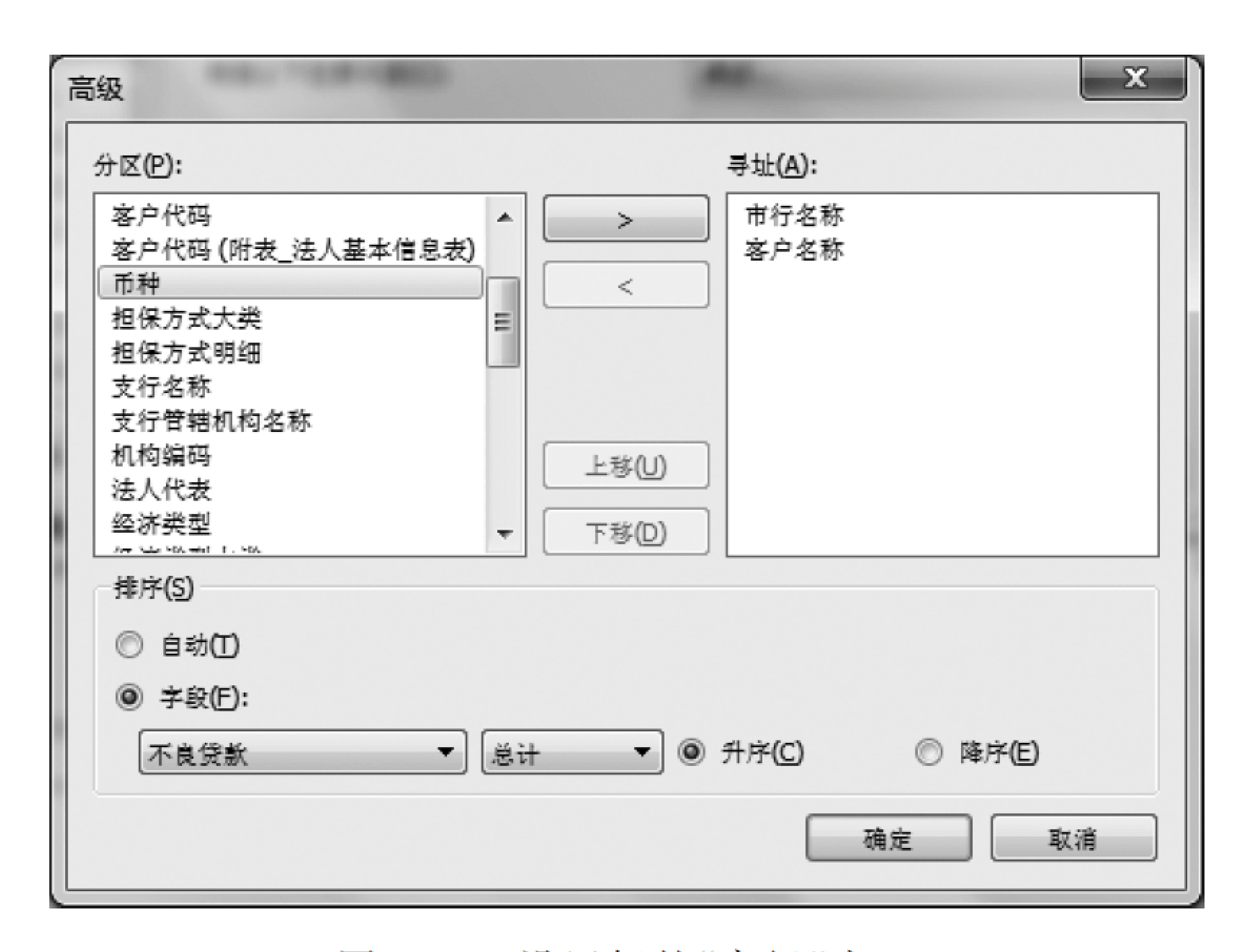


图 5-29 设置好的“高级”窗口

在“表计算”窗口中，在“所在级别”下拉列表框中选择“客户名称”，在“重新启动间隔”下拉列表框中选择“市行名称”。设置好的形式如图 5-30 所示。

单击“确定”按钮，关闭“表计算”窗口。

(3) 设置筛选条件

将“借款年份”拖放到“筛选器”，勾选“2008”，假设只分析 2008 年不良贷款情况。

将“经济类型大类名称”拖放到“筛选器”，勾选“集体企业”，这里只分析“集体企业”的不良贷款情况。

将“市行名称”拖放到“筛选器”，选择“显示筛选器”，将“市行名称”的筛选显示在分析



图 5-30　设置好的“表计算”窗口

视图中，并勾选“F 市分行”“G 市分行”和“H 市分行”。

将“行”功能区的“排名”拖放到“筛选器”，在弹出的“筛选器”窗口中，勾选前 5 个排名(假设我们只考察每个市行发放不良贷款最多的前 5 家企业)。

将分析视图中的数据按“不良贷款”降序排序，将“不良贷款”拖放到“标记”卡的“标签”上。

最终的分析结果如图 5-31 所示，从图中可看到只列出了每个市行发放不良贷款最多的 5 个客户名称，并且对每个市行按不良贷款额降序排序。

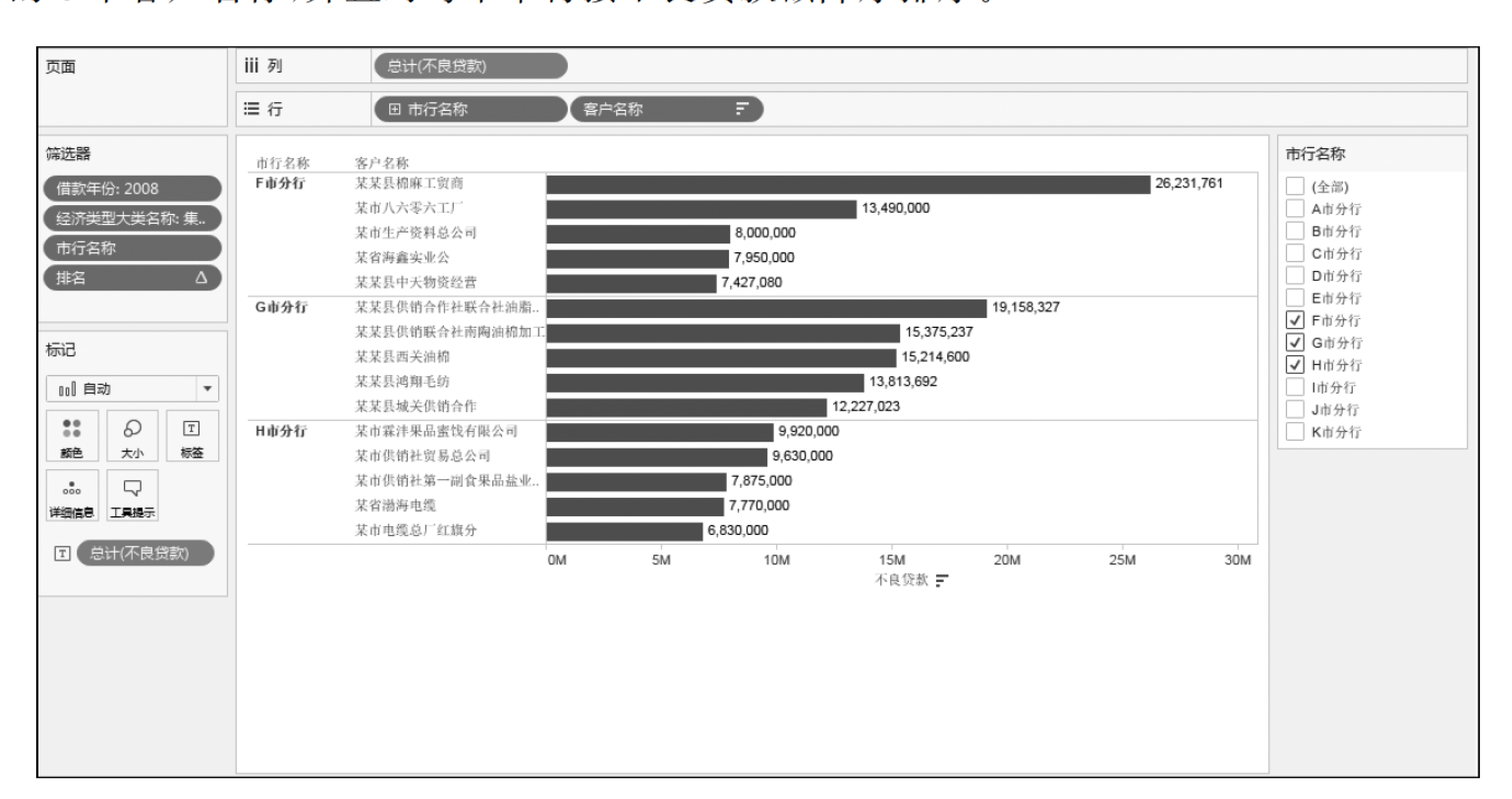


图 5-31　重点银行给重点客户发放不良贷款情况

5.1.4　各类贷款的不良贷款情况分析

1. 把握总体：各贷款类别在当年的不良贷款中的占比

分析目标：分析各个贷款大类的不良贷款情况，了解每个贷款大类在当年不良贷款中的占比。

分析实现过程：

(1) 设置分析视图

新建一个工作表。将“借款年份”拖放到“列”功能区,将“贷款类别大类”拖放到“行”功能区。

将“不良贷款”拖放到“颜色”和“文本”,将右端颜色设置为红色。

(2) 设置筛选条件

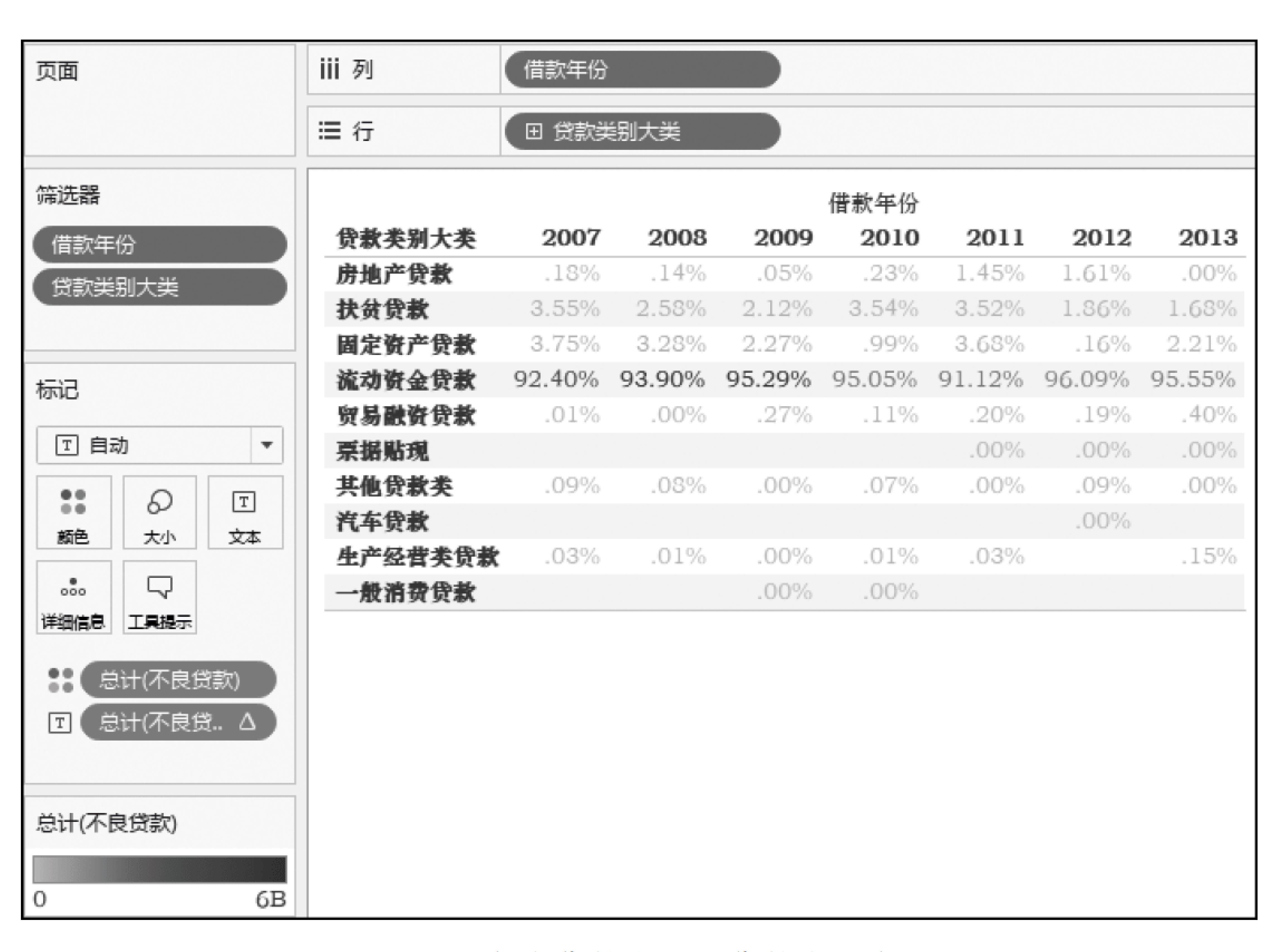
将“借款年份”拖放到“筛选器”,由于2007—2013年是不良贷款高峰年,因此,这里只勾选2007—2013年。

将“贷款类别大类”拖放到“筛选器”,去掉其中取值为“null”的贷款类别大类。

(3) 设置不良贷款占比

单击标签上的“总计(不良贷款)”的下三角按钮,在弹出的菜单中选择“快速表计算”→“总额百分比”。

产生的分析结果如图5-32所示,从图中可以看到每年90%以上的不良贷款都发放给了“流动资金贷款”,因此需要对该类贷款予以关注。



	借款年份						
贷款类别大类	2007	2008	2009	2010	2011	2012	2013
房地产贷款	.18%	.14%	.05%	.23%	1.45%	1.61%	.00%
扶贫贷款	3.55%	2.58%	2.12%	3.54%	3.52%	1.86%	1.68%
固定资产贷款	3.75%	3.28%	2.27%	.99%	3.68%	.16%	2.21%
流动资金贷款	92.40%	93.90%	95.29%	95.05%	91.12%	96.09%	95.55%
贸易融资贷款	.01%	.00%	.27%	.11%	.20%	.19%	.40%
票据贴现					.00%	.00%	.00%
其他贷款类	.09%	.08%	.00%	.07%	.00%	.09%	.00%
汽车贷款						.00%	
生产经营类贷款	.03%	.01%	.00%	.01%	.03%		.15%
一般消费贷款			.00%	.00%			

图5-32　各类贷款在不良贷款中的占比

2. 锁定重点:不良贷款率高的贷款细分类别

分析目标:找出不良贷款率高的贷款的细分类别。

新建一个工作表。将“贷款类别大类”拖放到“行”功能区,将“不良贷款率”拖放到“列”功能区和“标签”上,将“不良贷款率”的显示格式设置为百分比,将“不良贷款”拖放到“颜色”上。展开“行”功能区的“贷款类别大类”,产生的分析结果如图5-33所示。从该图可明显看出,虽然“一般企业流动资金贷款”的不良贷款额非常高(颜色越深数额越大),但其不良贷款率并不是非常高,为47.1%。而“自建房贷款”的不良贷款率最高,为100%。下面可进一步分析不良贷款率高的这些类型的贷款的具体情况。

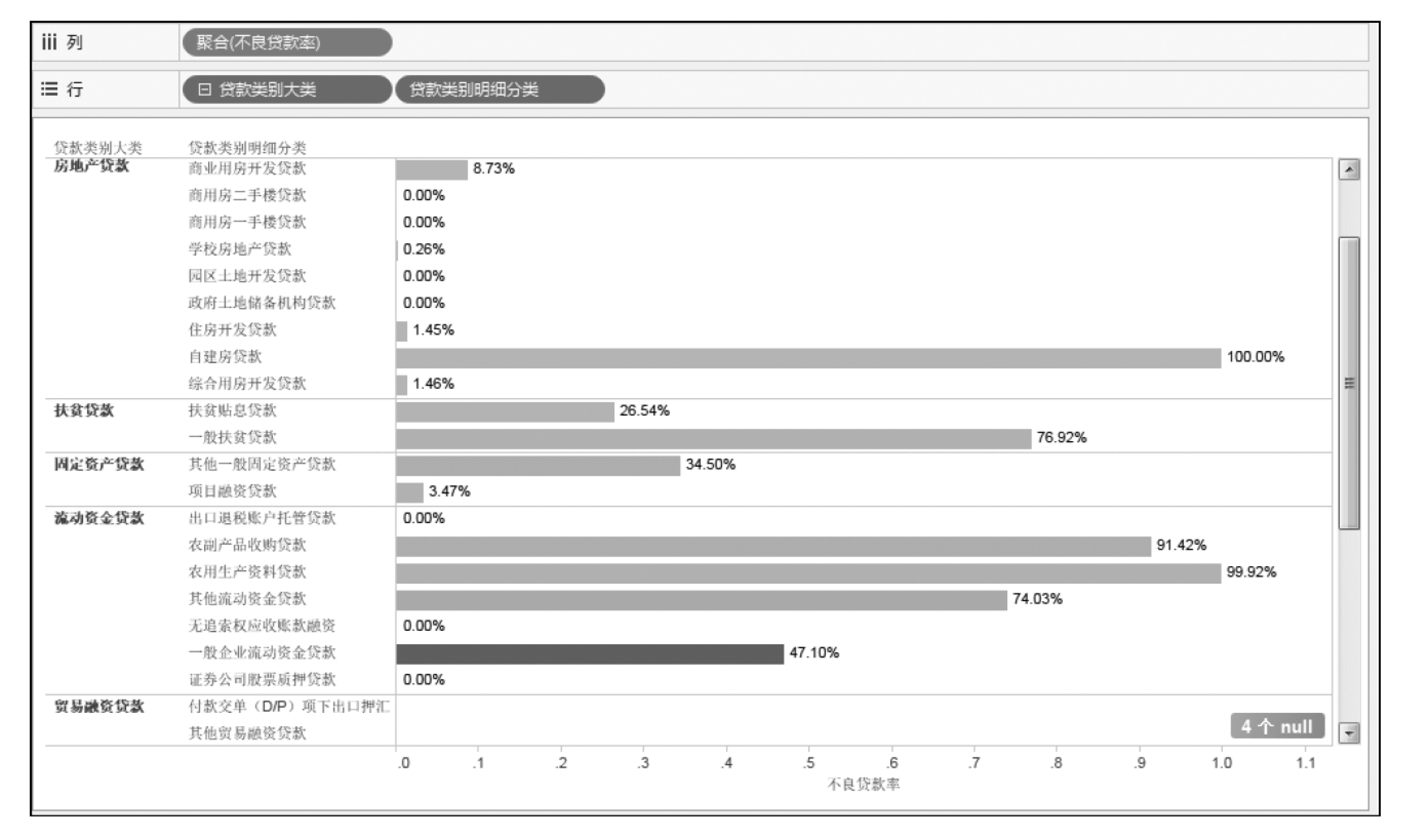


图 5-33　各贷款明细分类的不良贷款率

3. 深入分析：各银行发放不良贷款率高的贷款类别

分析目标：从图 5-33 的分析得知“自建房贷款”的不良贷款率最高，为 100%，但其不良贷款额很少，而且只有一笔贷款，因此，我们可以不对此类贷款进行分析。而“农用生产资料贷款”的不良贷款率非常高，为 99.92%，其不良贷款额也比较高，我们分析 2003—2011 年各银行给“农用生产资料贷款”发放的不良贷款数额。

分析实现过程：

(1) 设置分析视图

新建一个工作表。将“借款日期”拖放到“列”功能区，粒度选“年”，将“市行名称”拖放到“行”功能区。

将“不良贷款率”拖放到“文本”上。

(2) 设置筛选条件

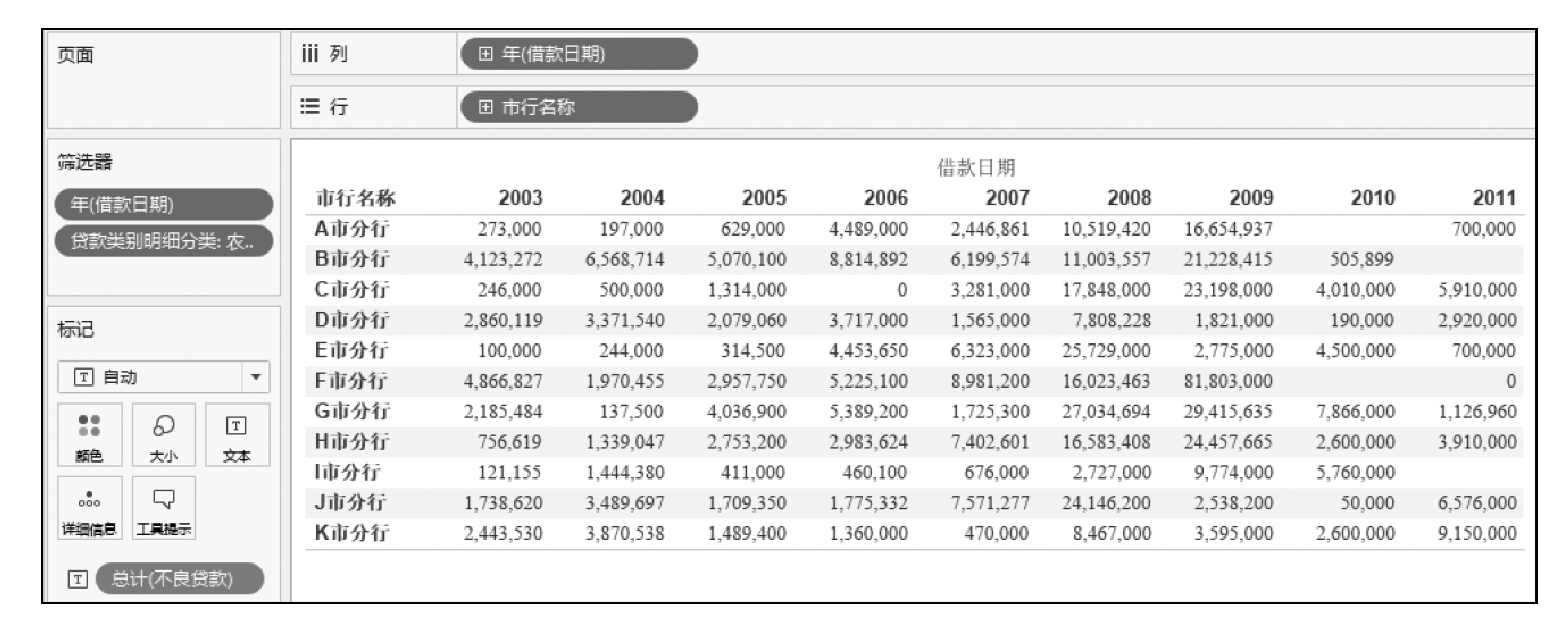
将“借款日期”拖放到“筛选器”，勾选 2003—2011 年的所有年。

将“贷款类别明细分类”拖放到“筛选器”，勾选“农用生产资料贷款”。

产生的分析结果如图 5-34 所示。这里用表格形式给出每个银行每年给“农用生产资料贷款”发放的不良贷款总额。

4. 关联分析：各银行发放不良贷款率高的贷款类别

分析目标：从之前的分析可知贷款类别为“农用生产资料贷款”，经济类型为“集体企业”的 2008 年和 2009 年的不良贷款额和不良贷款率都比较高。下面综合分析贷款类别为“农用生产资料贷款”、经济类型为“集体企业”的客户中，2008 年和 2009 年都有哪些银行给这些类别的客户中的哪些具体客户发放了比较多的不良贷款。



市行名称	2003	2004	2005	2006	2007	2008	2009	2010	2011
A市分行	273,000	197,000	629,000	4,489,000	2,446,861	10,519,420	16,654,937		700,000
B市分行	4,123,272	6,568,714	5,070,100	8,814,892	6,199,574	11,003,557	21,228,415	505,899	
C市分行	246,000	500,000	1,314,000	0	3,281,000	17,848,000	23,198,000	4,010,000	5,910,000
D市分行	2,860,119	3,371,540	2,079,060	3,717,000	1,565,000	7,808,228	1,821,000	190,000	2,920,000
E市分行	100,000	244,000	314,500	4,453,650	6,323,000	25,729,000	2,775,000	4,500,000	700,000
F市分行	4,866,827	1,970,455	2,957,750	5,225,100	8,981,200	16,023,463	81,803,000		0
G市分行	2,185,484	137,500	4,036,900	5,389,200	1,725,300	27,034,694	29,415,635	7,866,000	1,126,960
H市分行	756,619	1,339,047	2,753,200	2,983,624	7,402,601	16,583,408	24,457,665	2,600,000	3,910,000
I市分行	121,155	1,444,380	411,000	460,100	676,000	2,727,000	9,774,000	5,760,000	
J市分行	1,738,620	3,489,697	1,709,350	1,775,332	7,571,277	24,146,200	2,538,200	50,000	6,576,000
K市分行	2,443,530	3,870,538	1,489,400	1,360,000	470,000	8,467,000	3,595,000	2,600,000	9,150,000

图 5-34　各市行每年给“农用生产资料贷款”发放的不良贷款数额

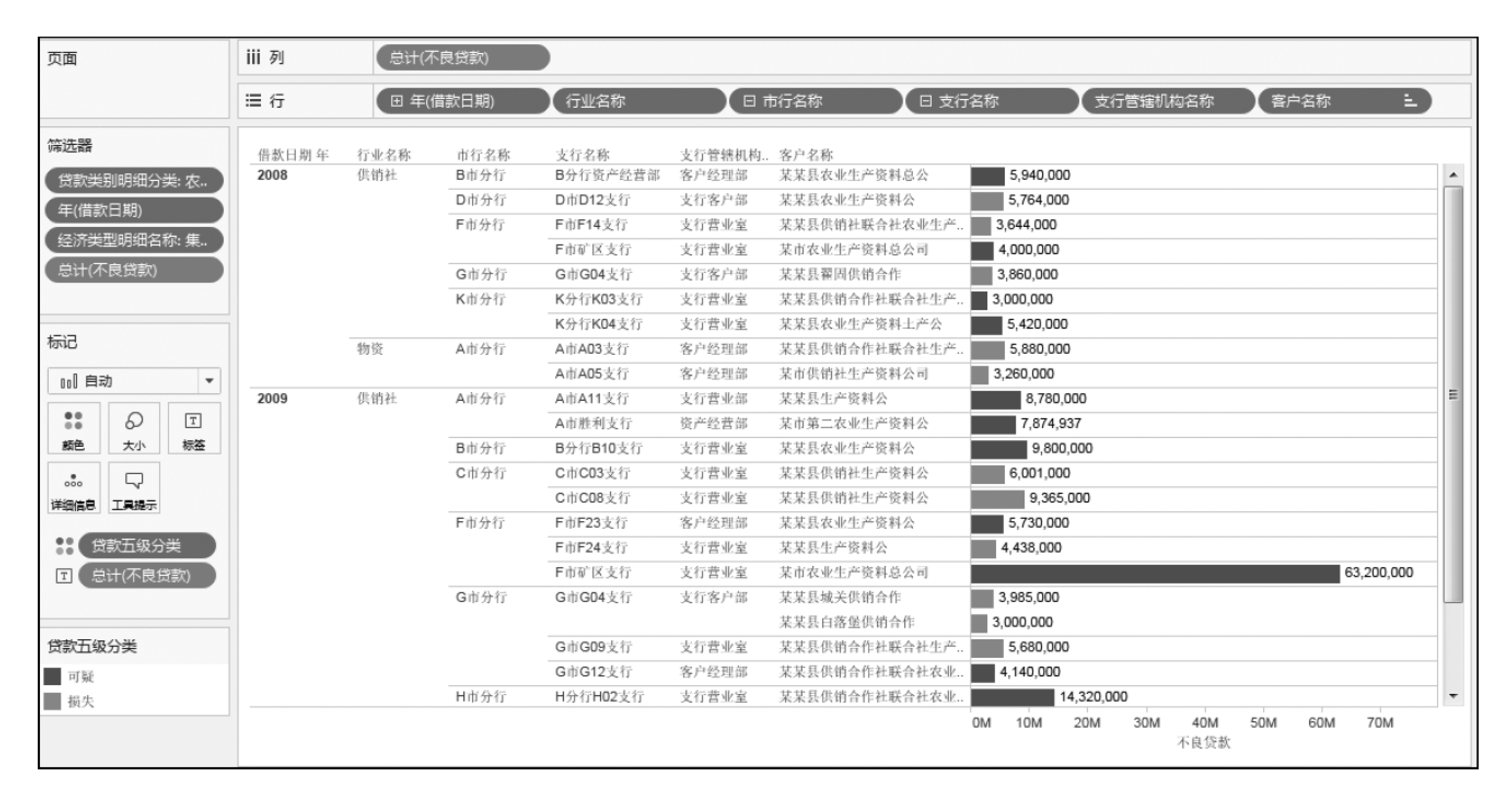
（1）设置分析视图

新建一个工作表。将“不良贷款”拖放到“列”功能区；将“借款日期”拖放到“行”功能区，选择粒度为“年”；将“行业名称”“市行名称”“客户名称”分别拖放到“行”功能区，并展开“市行名称”→“支行名称”→“支行管辖机构名称”；将“贷款五级分类”拖放到“颜色”；将“不良贷款”拖放到“标签”。

（2）设置分析条件

将“贷款类别明细分类”拖放到“筛选器”，并只勾选“农用生产资料贷款”；将“借款日期”拖放到“筛选器”，并勾选“2008”和“2009”；将“经济类型明细名称”拖放到“筛选器”，勾选“集体企业”；将“不良贷款”拖放到“筛选器”，并将筛选条件设为：＞＝3,000,000。

产生的分析结果如图 5-35 所示，从图中可以看到 2009 年“F 市矿区支行营业室”给“某市农业生产资料总公司”发放的“可疑”类不良贷款非常多。



借款日期 年	行业名称	市行名称	支行名称	支行管辖机构..	客户名称	不良贷款
2008	供销社	B市分行	B分行资产经营部	客户经理部	某某县农业生产资料总公	5,940,000
		D市分行	D市D12支行	支行客户部	某某县农业生产资料公	5,764,000
		F市分行	F市F14支行	支行营业室	某某县供销社联合社农业生产..	3,644,000
			F市矿区支行	支行营业室	某市农业生产资料总公司	4,000,000
		G市分行	G市G04支行	支行客户部	某某县瞿固供销合作	3,860,000
		K市分行	K分行K03支行	支行营业室	某某县供销合作社联合社生产..	3,000,000
			K分行K04支行	支行营业室	某某县农业生产资料上产公	5,420,000
	物资	A市分行	A市A03支行	客户经理部	某某县供销合作社联合社生产..	5,880,000
			A市A05支行	客户经理部	某市供销社生产资料公司	3,260,000
2009	供销社	A市分行	A市A11支行	支行营业部	某某县生产资料公	8,780,000
			A市胜利支行	资产经营部	某市第二农业生产资料公	7,874,937
		B市分行	B分行B10支行	支行营业室	某某县农业生产资料公	9,800,000
		C市分行	C市C03支行	支行营业室	某某县供销社生产资料公	6,001,000
			C市C08支行	支行营业室	某某县供销社生产资料公	9,365,000
		F市分行	F市F23支行	客户经理部	某某县农业生产资料公	5,730,000
			F市F24支行	支行营业室	某某县生产资料公	4,438,000
			F市矿区支行	支行营业室	某市农业生产资料总公司	63,200,000
		G市分行	G市G04支行	支行客户部	某某县城关供销合作	3,985,000
					某某县白落堡供销合作	3,000,000
			G市G09支行	支行营业室	某某县供销合作社联合社生产..	5,680,000
			G市G12支行	客户经理部	某某县供销合作社联合社农业..	4,140,000
		H市分行	H分行H02支行	支行营业室	某某县供销合作社联合社农业..	14,320,000

图 5-35　关联多种信息的不良贷款分析

至此，我们从多个角度对不良贷款的发放情况进行了分析，找出了不良贷款发放多的年份、发放不良贷款多的银行，同时找出了这些不良贷款都具体发放给了哪些企业或个人，从而为今后减少或避免不良贷款的发放、提高信贷资产质量，提供了可靠的参考依据。

5.2　保险公司客户索赔分析

在保险行业中常有大量的数据需要分析，时刻了解客户的索赔、公司的赔付情况，对保险公司来说非常重要。本案例以一家保险公司的索赔、赔付数据为例，分析该公司在全国各省各年龄段的不同性别的客户的索赔额及实际赔付额情况，并分析哪个或哪些省的情况比较异常以及索赔与赔付的趋势，同时演示了根据客户索赔的金额，预测实际赔付额会是多少。

该保险公司客户索赔分析数据存储在一个 Excel 文件中，结构为：

保险公司客户索赔数据(日期，客户性别，索赔单号，保险单号，区域，身份，城市，客户年龄，索赔额，赔付额)

5.2.1　数据准备

1. 连接数据源

在 Tableau 建立数据源窗口，选择连接到“Excel”，选择“保险公司客户索赔数据.xlsx”。此时连接数据源窗口如图 5-36 所示。

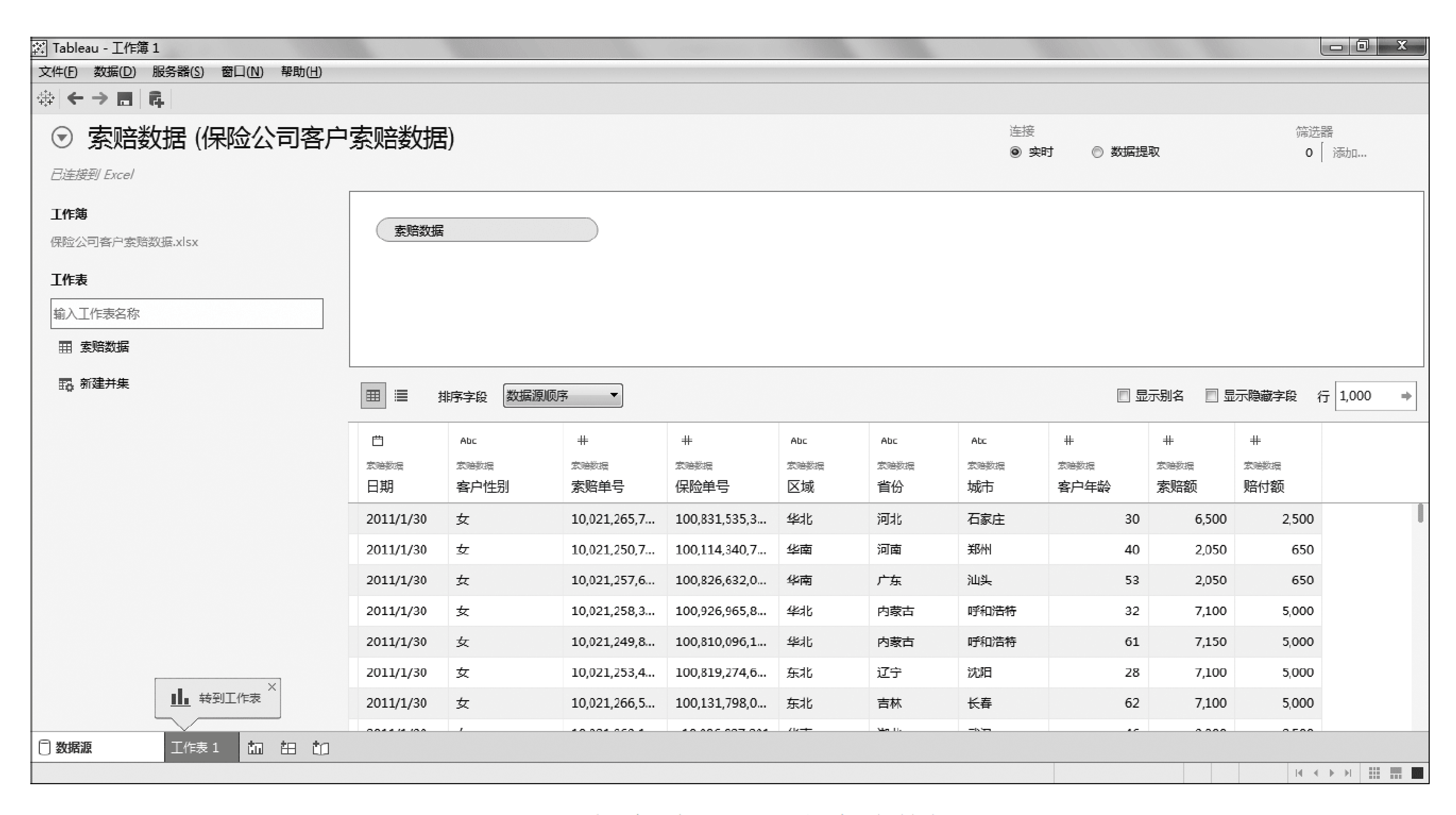


图 5-36　连接“保险公司客户索赔数据.xlsx”

2. 数据转换

在工作表的维度列表中，单击“省份”的下三角按钮，在弹出的菜单中选择“地理角色”→“省/市/自治区”。

在“度量”列表框中，单击“保险单号”的下三角按钮，在弹出的菜单中选择“转换为维度”。同样将“索赔单号”也转换为维度。

3. 定义计算字段

```
年龄区间 =
  IF [客户年龄] < 30 THEN '20-29'
  ELSEIF [客户年龄] < 40 THEN '30-39'
  ELSEIF [客户年龄] < 50 THEN '40-49'
  ELSEIF [客户年龄] < 60 THEN '50-59'
  ELSEIF [客户年龄] < 70 THEN '60-69'
  END
```

5.2.2 数据分析

1. 索赔分析

分析目标：分析各地区、各年龄段的客户的平均赔付金额，以及事故次数与平均赔付额之间的关系。

（1）定义计算字段

事故次数＝COUNT([索赔单号])

（2）设置分析视图

将“事故次数”“赔付额”分别拖放到“列”功能区和“行”功能区，并将“赔付额”的计算方式改为“平均值”。

将“省份”拖放到“标记”卡的“详细信息”上，将“客户性别”分别拖放到“形状”和“颜色”上。单击“标记”卡中“客户性别”形状的下三角按钮，选择“编辑形状”，在弹出的“编辑形状”对话框中，在“选择形状板”下拉列表框中选择“性别”，分别选中“男”和“女”的图标，如图 5-37 所示。

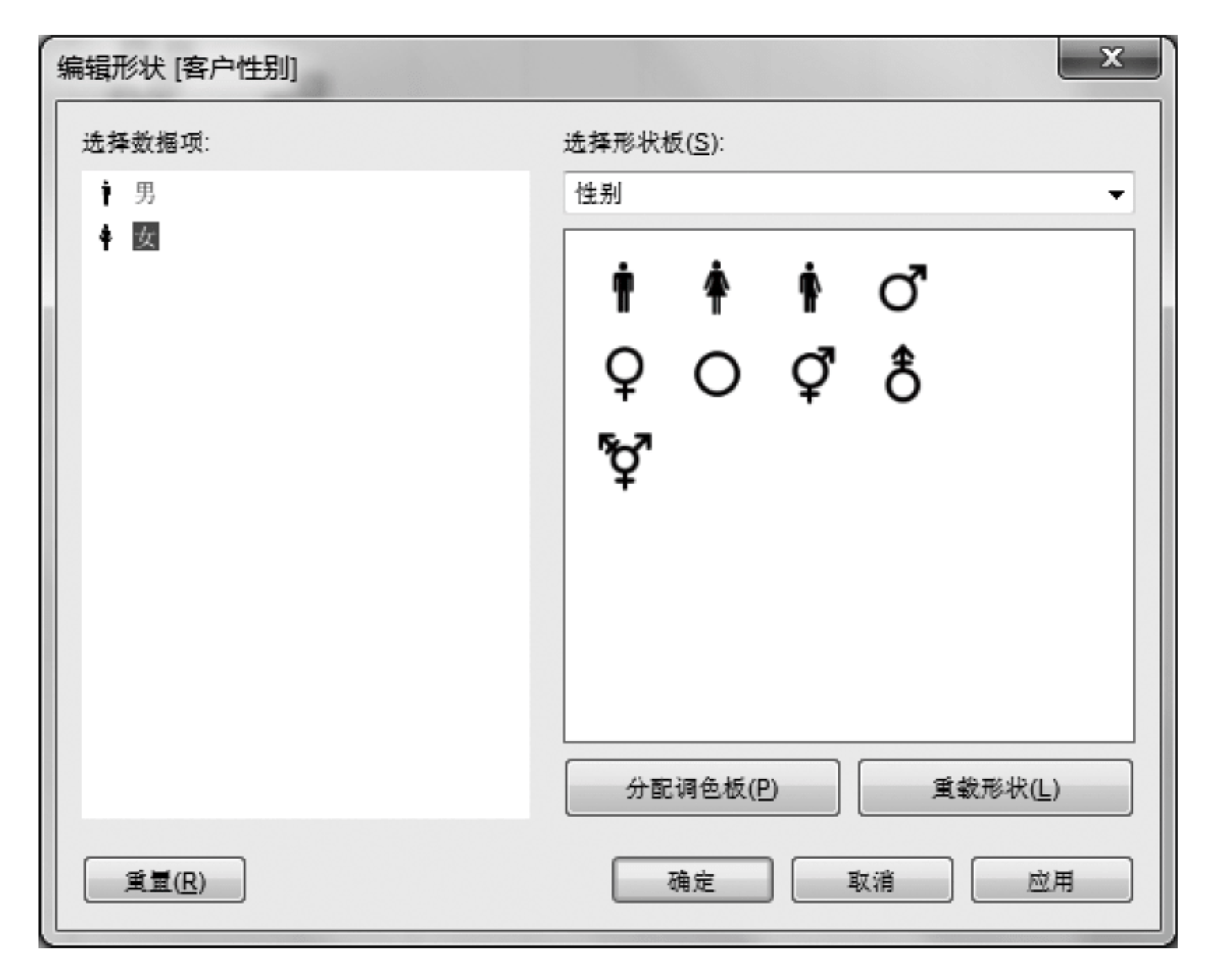


图 5-37　选择各性别的形状

至此生成的分析视图如图 5-38 所示。

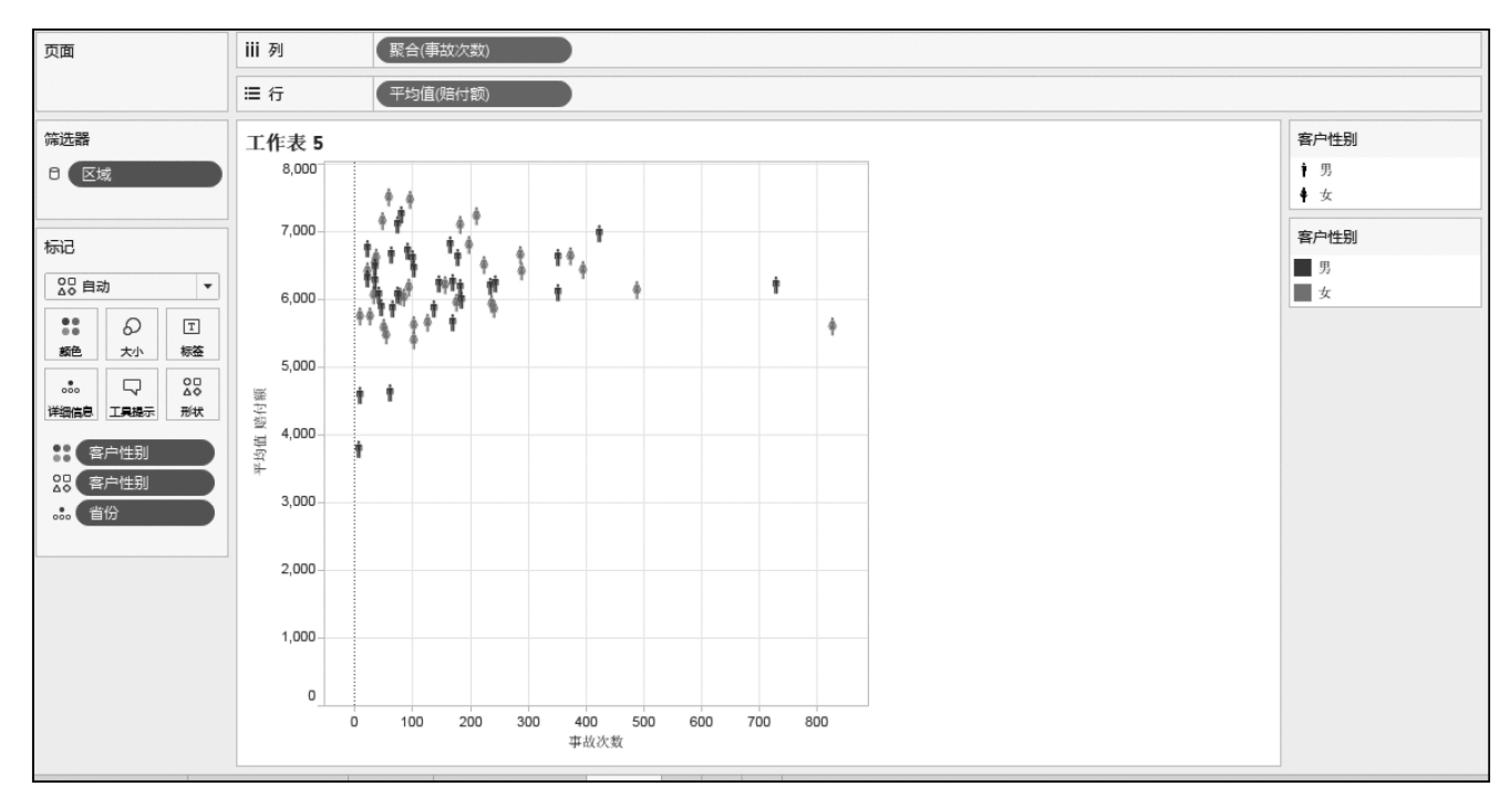


图 5-38　按客户性别显示的分析视图

(3) 添加参考线

右键单击横轴上的事故次数，在弹出的菜单中选择“添加参考线”，弹出“添加参考线、参考区间或框”的窗口，在此窗口中进行如下设置：

- 参考线形式选择“分布”;
- “范围”选择“每单元格”;
- “值”选择“标准差”，并将标准差因子设置为(−2,2);
- “标签”选择“无”;
- 在“格式”部分，“线”选择一种细线，“填充”选择浅灰色。

设置好后的形式如图 5-39 所示。单击“确定”按钮关闭此窗口。

右键单击纵轴，选择“编辑轴”，将纵轴标题改为“平均赔付额”。

然后再次右键单击纵轴，在弹出的菜单中选择“添加参考线”，在弹出的“添加参考线、参考区间或框”窗口中，进行同横轴参考线相同的设置。

至此，分析视图样式如图 5-40 所示。

(4) 添加详细信息

将“区域”“索赔额”“客户年龄”分别拖放到“标记”卡的“详细信息”上，并将“索赔额”“客户年龄”的计算方式改为“平均值”。

(5) 添加筛选条件

将“平均值(赔付额)”“平均值(索赔额)”“区域”“年龄区间”分别添加到“筛选器”中，并选择“显示筛选器”。将“平均值(赔付额)”和“平均值(索赔额)”的筛选条件选为“至少”，如图 5-41 所示。将“区域”和“年龄区间”筛选器显示格式均设置为“单值(下拉列表)”。

单击筛选条件区域“平均值(赔付额)”的下三角按钮，从弹出的菜单中选择“编辑标题”，将此标题改为“请选择赔付额区间”。同理将“平均值(索赔额)”“区域”“年龄区间”的



图 5-39　设置参考线格式

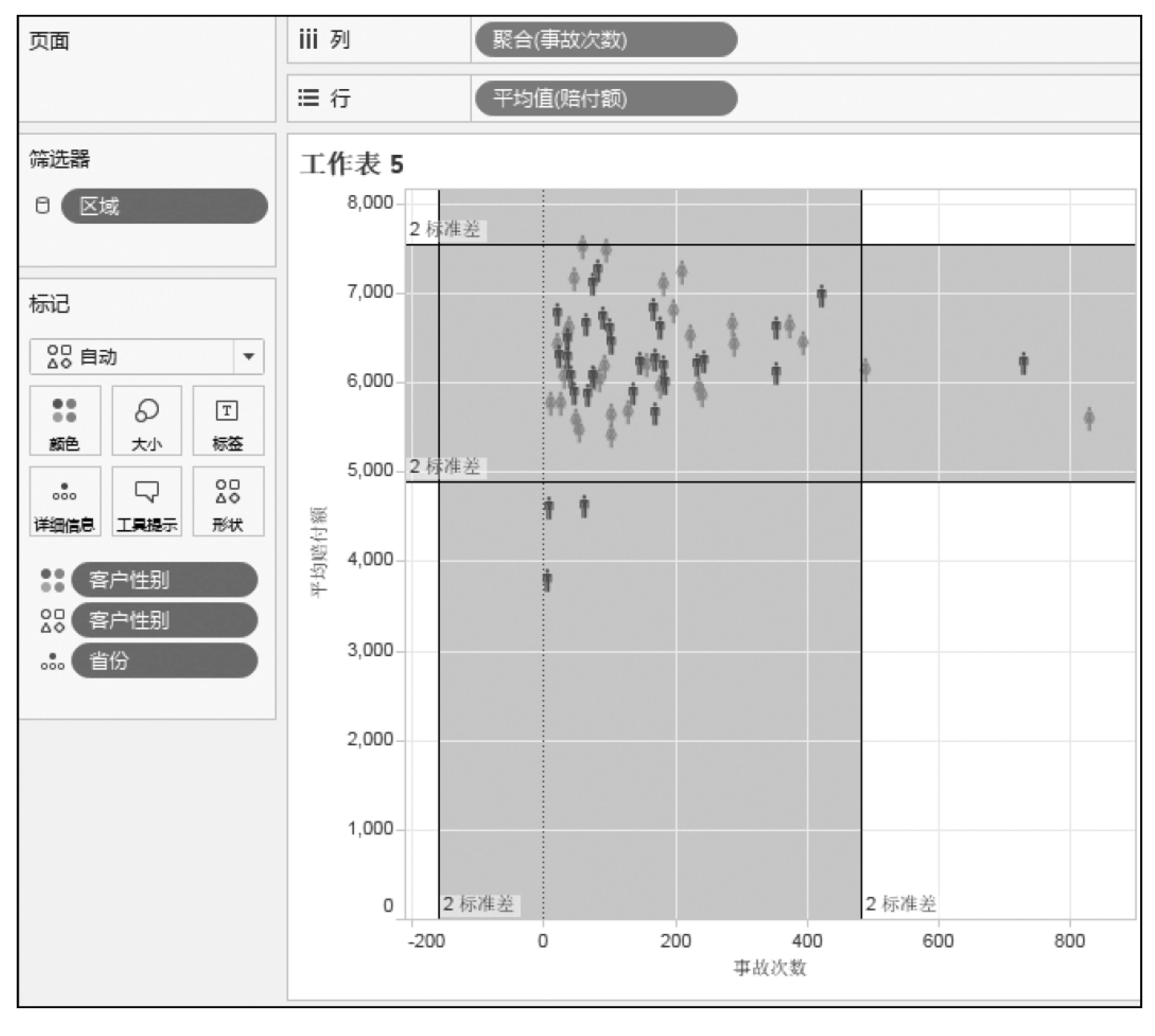


图 5-40　添加完参考线后的分析视图



图 5-41　将“平均赔付额”的筛选条件选为“至少”

筛选标题分别改为：“请选择索赔额区间”“请选择区域”“请选择年龄区间”。

至此，分析视图的样式如图 5-42 所示。

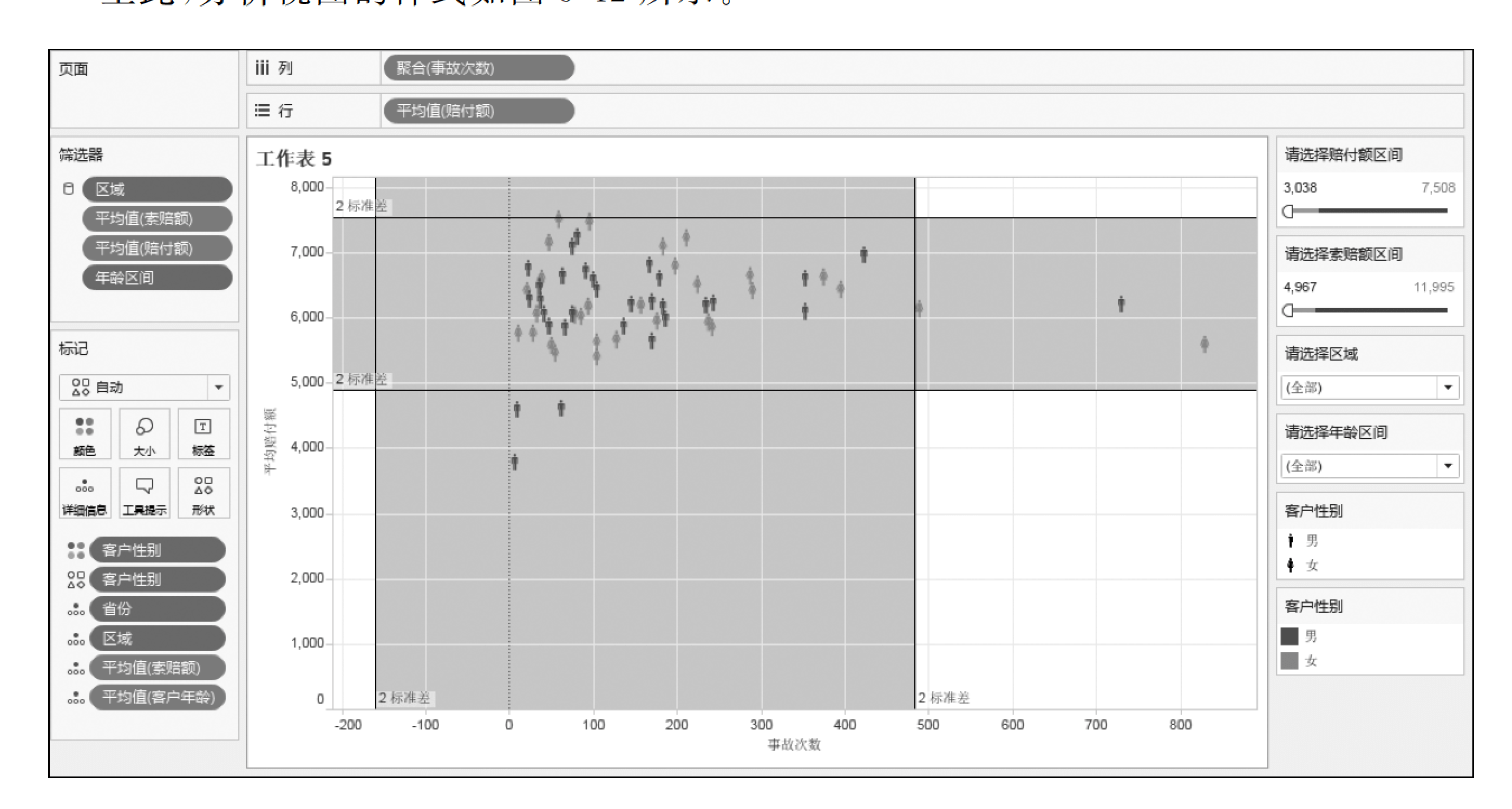


图 5-42　设置完筛选条件后的分析视图

(6) 添加趋势线

为了分析事故次数与平均赔付额之间的关系，可以采取添加趋势线的方法。实现方法如下：

右键单击分析视图中的任意位置，在弹出的菜单中选择“趋势线”→“显示趋势线”，分析视图效果如图 5-43 所示。

从图中可以看出，“直线”型的趋势线并不能很好地模拟事故次数与平均赔付额之间的关系，我们可以对趋势线类型进行修改，使趋势线能更好地体现事故次数与平均赔付额之间的关系。修改方法如下：

选中任意一条趋势线，在随之出现的选项中选择“编辑”（或者在趋势线上右击鼠标，在弹出的菜单中选择“编辑趋势线”），弹出“趋势线选项”窗口，在此窗口中进行如下设置：

- 在“模型类型”部分勾选“多项式”，并将度设置为 2；
- 在“选项”部分的“包括以下字段作为因素”列表框中确认选中“客户性别”；
- 取消对“允许按颜色绘制趋势线”等选项的勾选，只显示总体趋势线即可。

图 5-43　添加趋势线后的分析视图

趋势线的最终设置如图 5-44 所示。修改好趋势线后的分析视图如图 5-45 所示。



图 5-44　设置趋势线选项

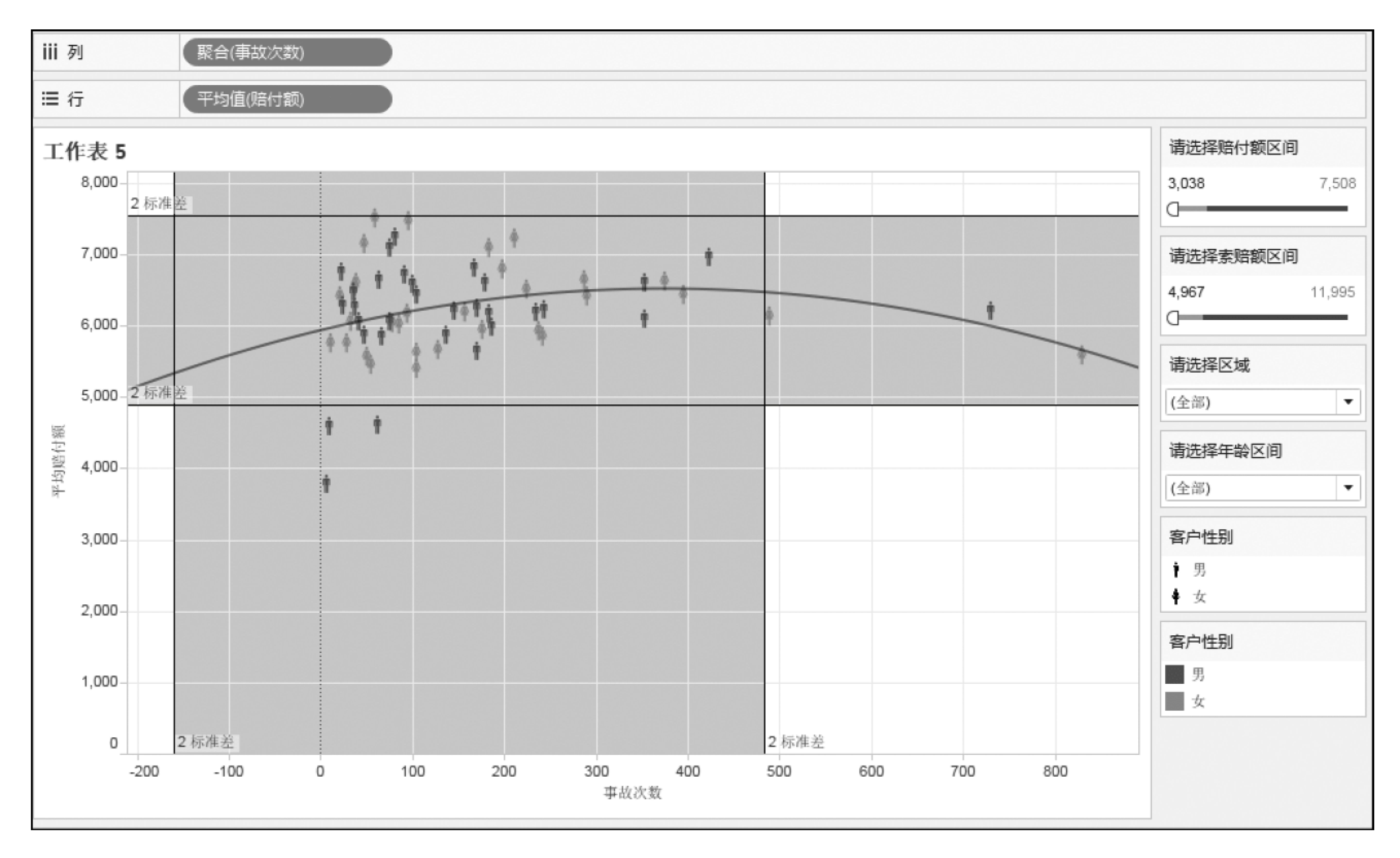


图 5-45　修改好趋势线类型后的分析视图

从图 5-45 可以很容易发现事故次数或平均赔付额在两个标准差之外的异常点，另外，可以通过使用视图右侧区域的筛选器方便地钻取到某一层数据，例如，将“区域”选为“华北”，年龄段选为“50～59”，则视图立即变为如图 5-46 所示的样式，从该图很容易发现，在华北地区，年龄为 50～59 的人群中，事故次数和平均赔付额基本都算正常。

图 5-46　“华北”地区 50～59 年龄段的赔偿的分析视图

如果将“区域”改为“华东”，则分析视图如图 5-47 所示。从图中可看到有少量事故

次数和平均赔付额不太正常。

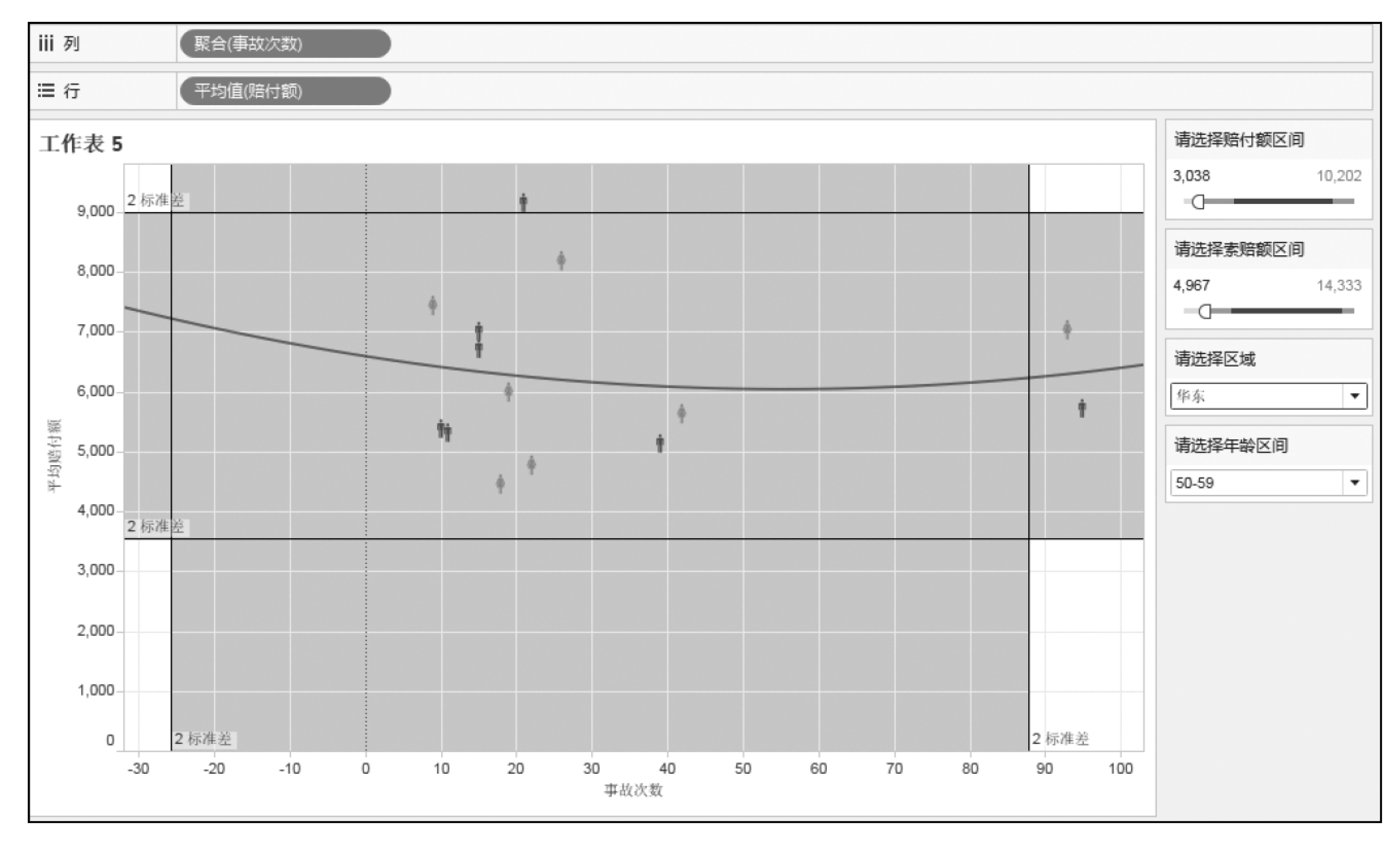


图 5-47 “华东”地区 50～59 年龄段的赔偿的分析视图

至此完成了索赔分析，将此工作表命名为“索赔分析”。

2. 各省的索赔额与赔付额趋势

分析目标：观察各省的客户索赔额与赔付额之间的趋势关系。

(1) 设置分析视图

- 新建一个工作表，按住 Ctrl 键，依次选中“省份”“赔付额”“索赔额”，单击“智能显示”，选择其中推荐的“散点图”。
- 单击工具栏中的“交换”图标，交换横、纵坐标轴。
- 单击“分析”菜单，取消对“聚合度量”的勾选。
- 右键单击视图中的任意位置，在弹出的菜单中选择“趋势线”→“显示趋势线”。

至此，生成的分析视图如图 5-48 所示。从图中可以看到索赔额与赔付额之间是线性关系，选择某个省，则显示该省的索赔额与赔付额之间的线性方程，通过该线性方程，可以预测，当某个省某个客户索赔一定金额时，最后可能需要赔付多少金额。

(2) 将工作表命名为“各省索赔与赔付情况”

3. 各省索赔额与赔付额排序

分析目标：哪个省的索赔额最多、哪个省的实际赔付额最多，以及每个省的索赔额与赔付额之间的差异情况。

新建一个工作表，将“省份”拖放到“行”功能区，将“索赔额”“赔付额”分别拖放到“列”功能区。

右键单击横轴上的“索赔额”，在弹出的菜单中选择“编辑轴”，弹出如图 5-48 所示的“编辑轴”窗口。在“范围”部分选中“固定”，“固定开始”值设为 0，“固定结束”值设为

16,000,000。在“比例”区域，勾选“倒序”(如图 5-49 所示)，然后单击“确定”按钮关闭此窗口。为保证横轴上的“赔付额”的坐标轴刻度与“索赔额”一致，右键单击横轴上的“赔付额”，选择“编辑轴”，进行如图 5-50 所示的设置。

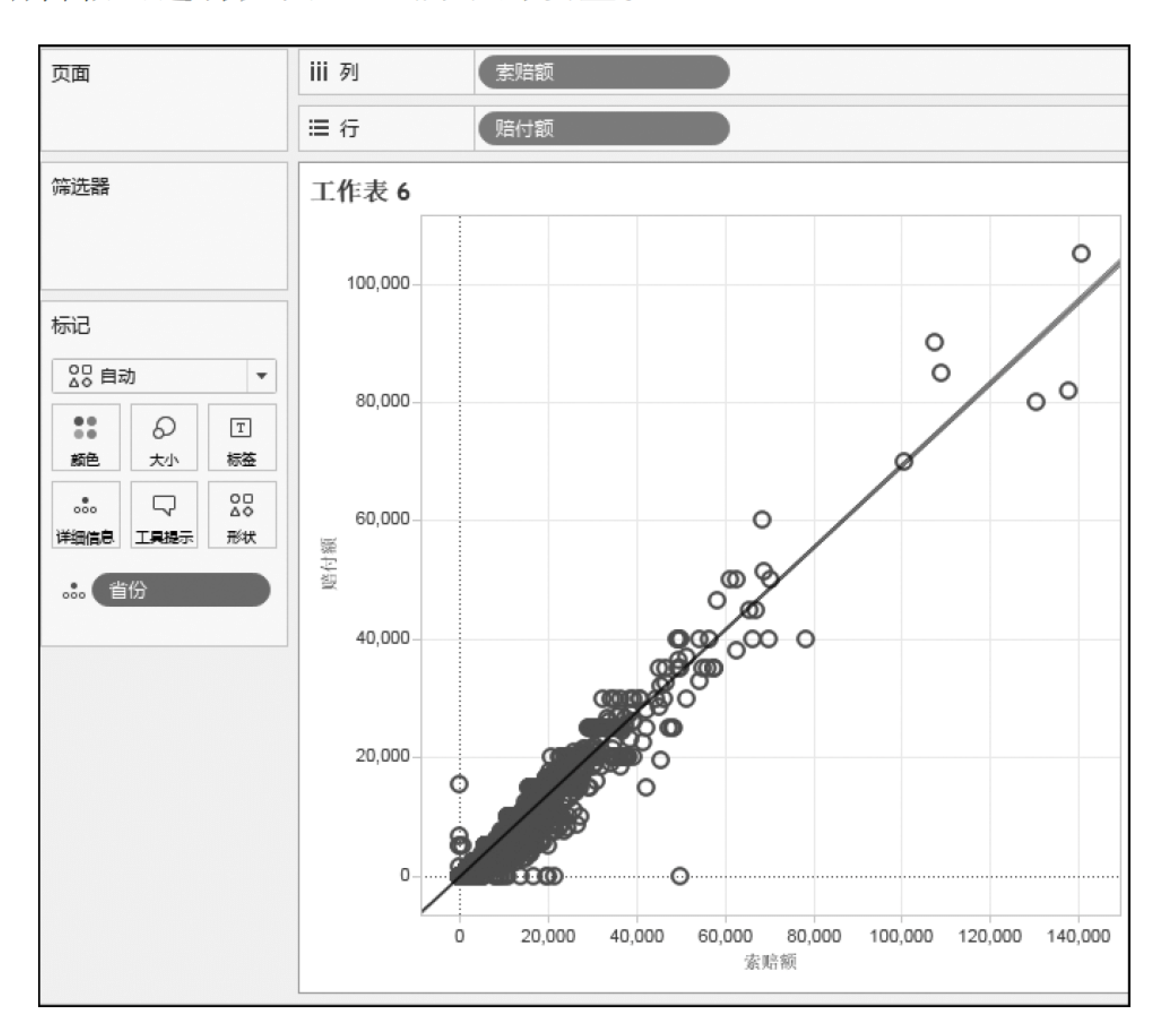


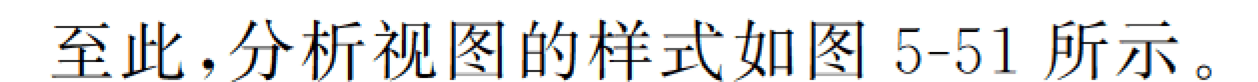
图 5-48　各省索赔额与赔付额趋势



图 5-49　“索赔额”的“编辑轴”窗口



图 5-50 “赔付额”的“编辑轴”设置


至此,分析视图的样式如图 5-51 所示。

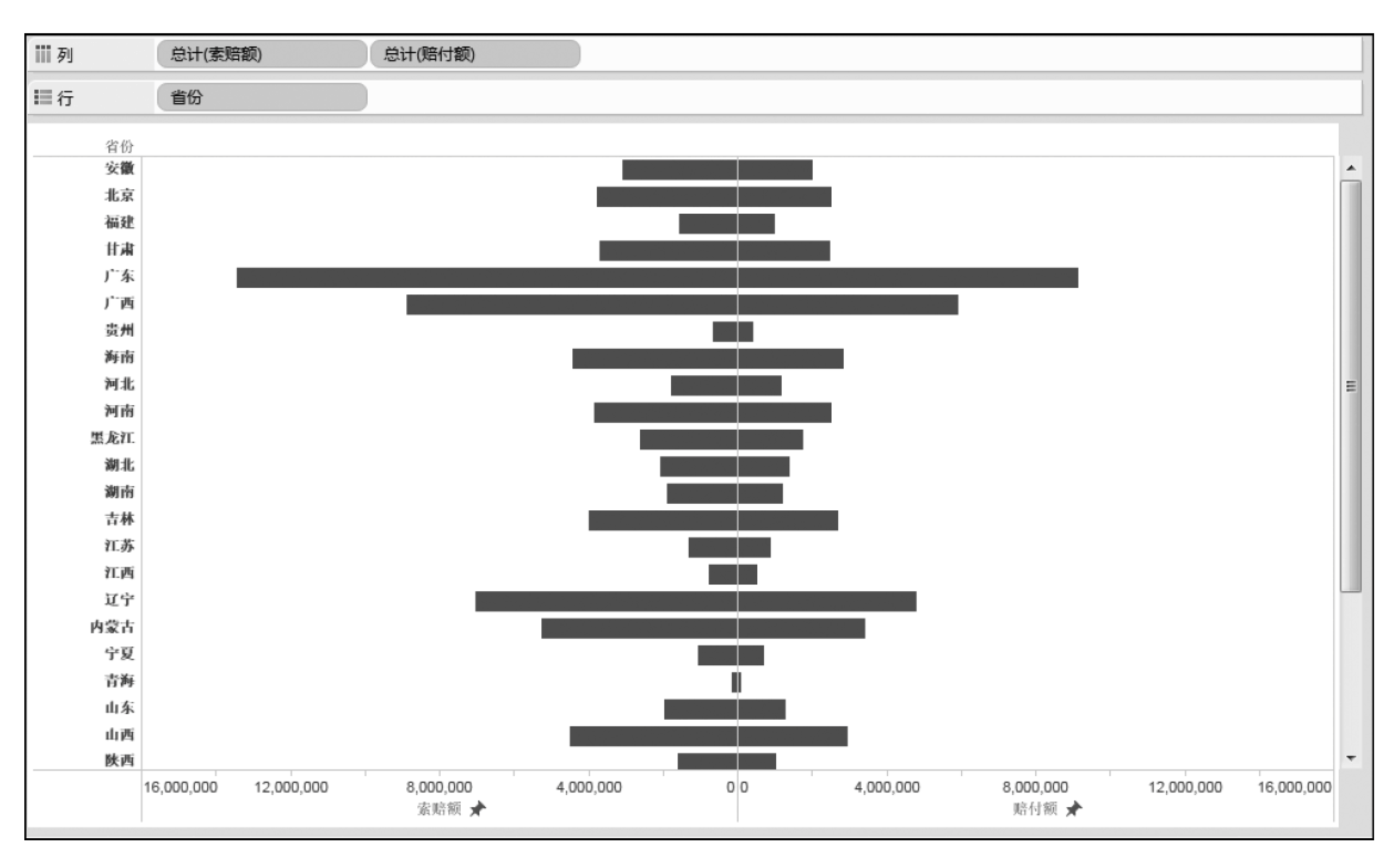


图 5-51 同步两个横轴刻度后的分析视图

选中“标记”中的“总计(赔付额)”,单击“颜色”,将“总计(赔付额)”设置为红色。

单击工具栏中的“降序”图标,将分析结果按“索赔额”降序排序。

单击工具栏中的“显示标记标签”图标,在分析视图中显示出具体数据值。

最终的分析视图如图 5-52 所示。将该工作表命名为“索赔与赔付降序分析”。

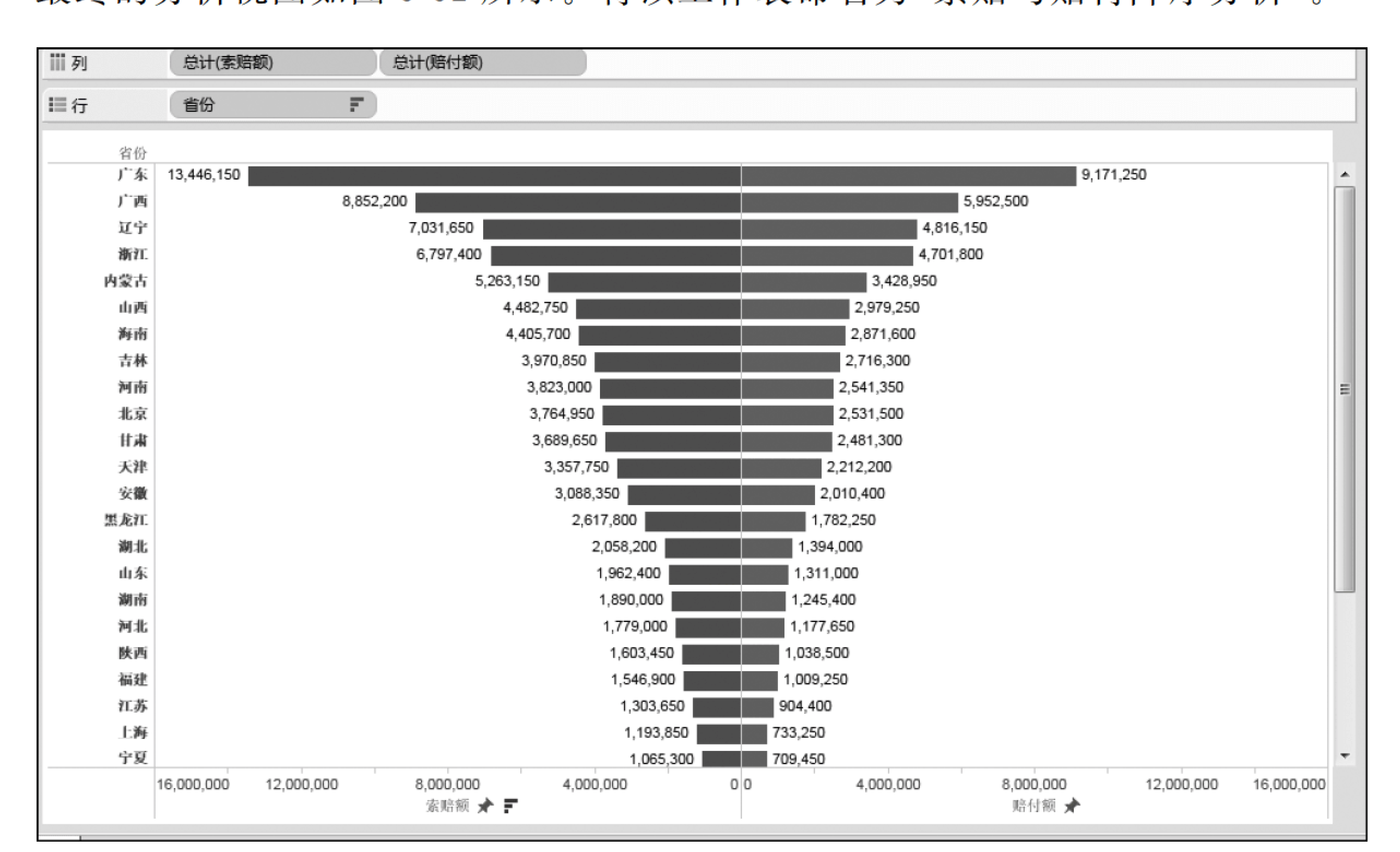


图 5-52　索赔与赔付降序分析视图

4. 综合分析

分析目标：组合已生成的各工作表，进行多角度的索赔与赔付分析。

(1) 设置仪表板布局

新建一个仪表板，将“索赔分析”“各省索赔与赔偿情况”“索赔与赔偿降序分析”工作表分别拖放到仪表板的合适位置。仪表板布局如图 5-53 所示。

(2) 美化仪表板外观

修改各工作表标题的显示方式，使仪表板界面更加美观。双击“索赔分析”工作表，弹出“编辑标题”窗口，在此窗口中，选中“＜工作表名称＞”，然后将颜色选为蓝色，显示方式为“居中”，如图 5-54 所示。

用同样的方法设置“各省索赔与赔偿情况”和“索赔与赔偿降序分析”工作表的标题。

设置仪表板的标题，方法如下：单击菜单栏上的“仪表板”，在出现的菜单项中勾选“显示标题”，然后在出现的仪表板标题上双击鼠标，弹出“编辑标题”窗口，在该窗口中将仪表板标题设为“索赔分析与预测”，字体居中显示。设置好的仪表板标题如图 5-55 所示。

至此仪表板的样式如图 5-56 所示。

(3) 设置筛选条件的使用范围

单击筛选条件中的“请选择区域”的下三角按钮，在弹出的菜单中选择“应用于工作表”→“选定的工作表”，在弹出的“将筛选器应用于工作表”窗口中，单击“仪表板上的所有项”按钮，勾选该仪表板上的全部工作表，如图 5-57 所示。用同样的方法设置“请选择年龄区间”筛选条件。

索赔分析

各省索赔与赔付情况

请选择赔付额区间 从 3,038

请选择索赔额区间 从 4,967

请选择区域 全部

请选择年龄区间 全部

索赔与赔付降序分析

省份	索赔额	赔付额
广东	13,446,150	9,171,250
广西	8,852,200	5,952,500
辽宁	7,031,650	4,816,150
浙江	6,797,400	4,701,800
内蒙古	5,263,150	3,428,950
山西	4,482,750	2,979,250
海南	4,405,700	2,871,600
吉林	3,970,850	2,716,300
河南	3,823,000	2,541,350
北京	3,764,950	2,531,500
甘肃	3,689,650	2,481,300
天津	3,357,750	2,212,200
安徽	3,088,350	2,010,400
黑龙江	2,617,800	1,782,250
湖北	2,058,200	1,394,000

图 5-53　初步建立的仪表板



图 5-54　设置工作表的标题显示格式

(4)设置筛选动作

创建一个筛选动作,当点击“索赔分析”中的某个点时,“各省索赔与赔偿情况”显示对应省份的数据,否则显示全部数据。

设置方法如下:

- 单击菜单栏的“仪表板”,选择“操作”,弹出“操作”窗口。



图 5-55　设置仪表板的标题显示格式

图 5-56　设置工作表标题显示格式

- 在“操作”窗口中单击“添加操作”按钮，选择“筛选器”，弹出“添加筛选器操作”窗口。
- 在该窗口中，在“名称”部分将该动作命名为“索赔与赔付筛选”。在“源工作表”中勾选“索赔分析”，在“运行操作方式”部分选择“选择”。在“目标工作表”部分勾选“各省索赔与赔付情况”，在“清除选定内容将会”部分选中“显示所有值”，表示当不选中任何内容时，显示全部值。设置结果如图 5-58 所示。



图 5-57　设置筛选条件的作用范围

• 单击“确定”按钮，关闭“添加筛选器操作”窗口。



图 5-58　设置仪表板的筛选动作

（5）设置突出显示

创建突出显示，每当在“索赔分析”中点击一个客户时，在“各省索赔与赔付降序分析”中都将突出显示对应省份的信息。或者当在筛选条件中指定某个区域时，在“各省索赔与赔付降序分析”中只突出显示该区域包含的省份的数据。

设置方法如下：

- 单击菜单上的“仪表板”，选择“操作”，弹出“操作”窗口。
- 在“操作”窗口中单击“添加操作”按钮，选择“突出显示”，弹出“添加突出显示操作”窗口。
- 在该窗口中，在“名称”部分将该动作命名为“突出显示各省”。在“源工作表”中勾选“索赔分析”，在“运行操作方式”部分选择“选择”。在“目标工作表”部分勾选“索赔与赔付降序分析” 和“索赔分析”。设置结果如图 5-59 所示。
- 单击“确定”按钮，关闭“添加突出显示操作”窗口。



图 5-59　设置仪表板的突出显示动作

至此，完成了仪表板的制作，将此仪表板命名为“索赔与预测分析”。

点击仪表板中“索赔分析”工作表中的某个图标，则“各省索赔与赔付情况”将同步变化为只显示所选定的省份的分析情况，并且在“索赔与赔付降序分析”工作表上将高亮度显示出该省的数据，如图 5-60 所示。

在仪表板中，将“请选择区域”的筛选条件选定为“华南”，将“请选择年龄区间”的筛选条件选定为“40-49”，仪表板的分析样式如图 5-61 所示。

图 5-60　在“索赔分析”中选中某个客户后的仪表板样式

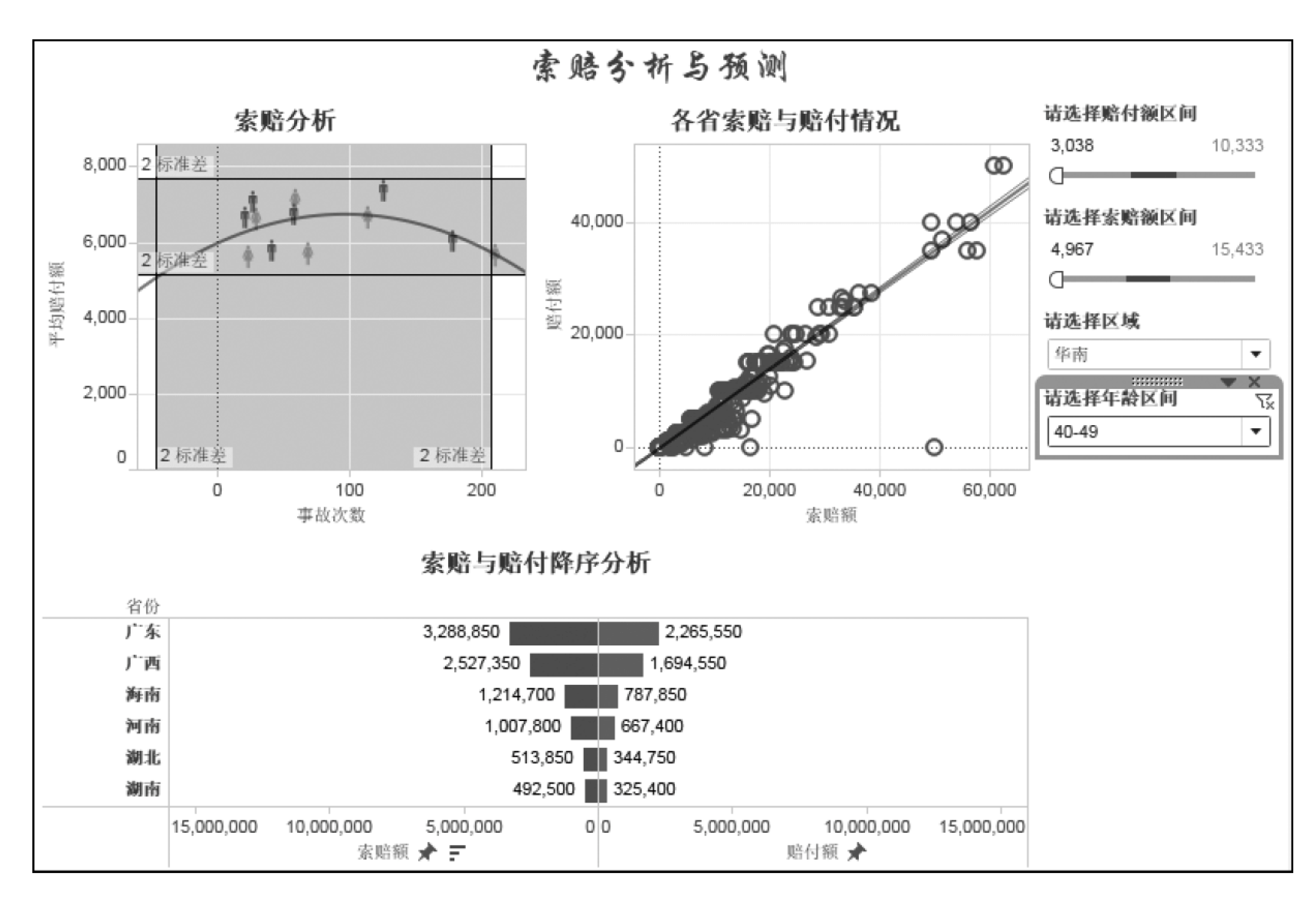


图 5-61　设定筛选条件后的仪表板样式

参考文献

[1] 刘汝焯. 审计分析模型算法：第 2 版[M]. 北京：清华大学出版社，2016.
[2] 沈浩，等. Tableau：数据可视化之极速 BI[M]. 北京：博易智讯，[出版年不详].
[3] 刘红阁，等. 人人都是数据分析师：Tableau 应用实战[M]. 北京：人民邮电出版社，2015.

第 6 章　可视化挖掘分析

本章通过两个案例演示用 RapidMiner 对数据进行可视化挖掘分析的过程，第一个案例以某航空公司油料成本支出审计为例，第二个案例为挖掘分析在推荐系统中的应用。我们通过这两个案例演示挖掘分析在不同领域的应用。

RapidMiner 是数据挖掘和商业预测分析领域的备受用户青睐的软件。RapidMiner 具有功能强大、用户入门快、操作简单的特点。用户使用它不需要任何编程知识，只需通过鼠标拖放，就能完成数据挖掘和分析的功能。

6.1　挖掘分析在审计线索特征发现中的应用

本节以某航空公司油料成本支出审计案例为例，介绍如何将 RapidMiner 用于特征发现。

6.1.1　案例背景

审计组在对某航空公司 2014 年度财务收支情况的审计中了解到，该公司 2014 年航油支出占总支出的 28.93%。因此，审计组决定将航空油料支出的真实性和合法性作为审计的一个重点内容。审计人员了解到，由于当时国内航油价格较国外航油价格高，国内部分航空公司存在利用飞机从国外带油的现象。为了解该公司是否确实存在这种现象，审计人员需要对其以往的航班记录进行检查，核实其加油量是否大于消耗量。

6.1.2　数据准备

审计人员采集了该公司航班生产管理系统的底层数据，其中包含了全部与航班飞行相关的信息。经过数据清理和验证，形成了三张审计中间表：分析表_飞行任务书、附表_航段信息表和附表_机型说明表。这些数据表的结构分别见表 6-1、表 6-2 和表 6-3。

表 6-1　分析表_飞行任务书

序号	字段名称	数据类型
1	记录号	整型
2	日期	日期
3	航班号	字符串
4	公司	字符串

续表

序号	字段名称	数据类型
5	飞行队	字符串
6	机号	字符串
7	机型	字符串
8	航班性质	字符串
9	航线分类	字符串
10	航段距离	整型
11	原存油	整型
12	新加油	整型
13	留存油	整型
14	加耗油差额	整型

表 6-2　附表_航段信息表

序号	字段名称	数据类型
1	航段	字符串
2	起飞地简码	字符串
3	起飞地航站名	字符串
4	起飞地航站类别	字符串
5	目的地简码	字符串
6	目的地航站名	字符串
7	目的地航站类别	字符串
8	航班类别	字符串

表 6-3　附表_机型说明表

序号	字段名称	数据类型
1	机型	字符串
2	全称	字符串

审计人员需要掌握加油量与耗油量差额，因此在“分析表_飞行任务书”表中增加了一个新的字段“加耗油差额”，其计算公式为“加耗油差额”＝“加油量”－“耗油量”。加油量由“新加油”字段反映，耗油量的计算公式为“原存油”＋“新加油”－“留存油”，因此，上述计算“加耗油差额”的公式可以进一步简化为“留存油”－“原存油”。

6.1.3 聚类分析

本案例采用数据挖掘中的聚类分析方法来寻找和发现审计线索。

聚类是根据“物以类聚”的原理，把一组数据对象划分成不同分类的过程(这样的一个分类又称为簇)，使同一簇中的数据对象有很大的相似性，而不同簇的数据对象之间则有很大的差异性。聚类分析是按照某种相似性自动聚合数据集，因此聚类结果不仅可以揭示数据间的内在联系与区别，还可以为进一步的数据分析与知识发现提供重要依据。聚类分析是数据挖掘技术中的重要组成部分，已经在教育、交通、医学、科学研究等领域获得了广泛应用。常用的聚类算法包括 K-Means、K-Medoids、DBSCAN、BIRCH、CURE、CHAMELEON 等，本案例采用 K-Means 聚类算法。

本案例中某航空公司燃油分析的审计数据表保存在 SQL Server 2012 中，聚类分析的软件是 RapidMiner 7.2。下面介绍从 SQL Server 2012 中提取审计数据，在 RapidMiner 中完成聚类分析的过程。

(1) 启动 RapidMiner 软件，单击“Repository”标签下方的“Add Data”按钮，弹出加载数据库的窗口，如图 6-1 所示。



图 6-1　加载数据库窗口

(2) 在如图 6-1 所示的图中选择“Database”选项，软件提示输入数据库所在的位置，如图 6-2 所示。单击“New Connection”按钮，软件提示输入与 SQL Server 数据库连接的有关参数，如图 6-3 所示。

(3) 数据库连接参数的详细介绍如表 6-4 所示。输入连接参数后，单击“OK”按钮。如果能正确连接到数据库，则软件显示数据库包含的全部数据表，用户可以进一步从这些数据表中查询提取所需要的数据，如图 6-4 所示。

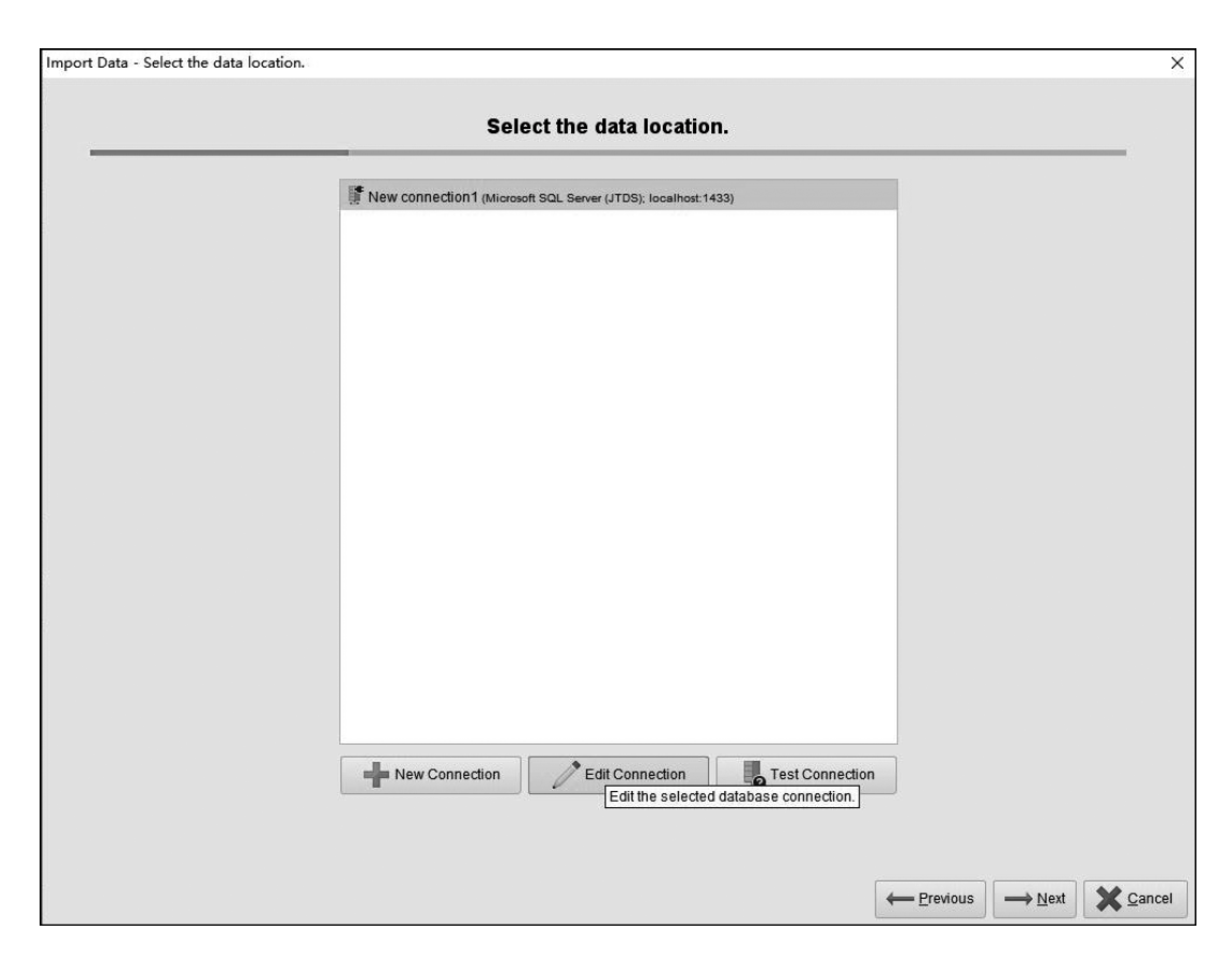


图 6-2　填写数据库所在位置



图 6-3　配置数据库连接参数

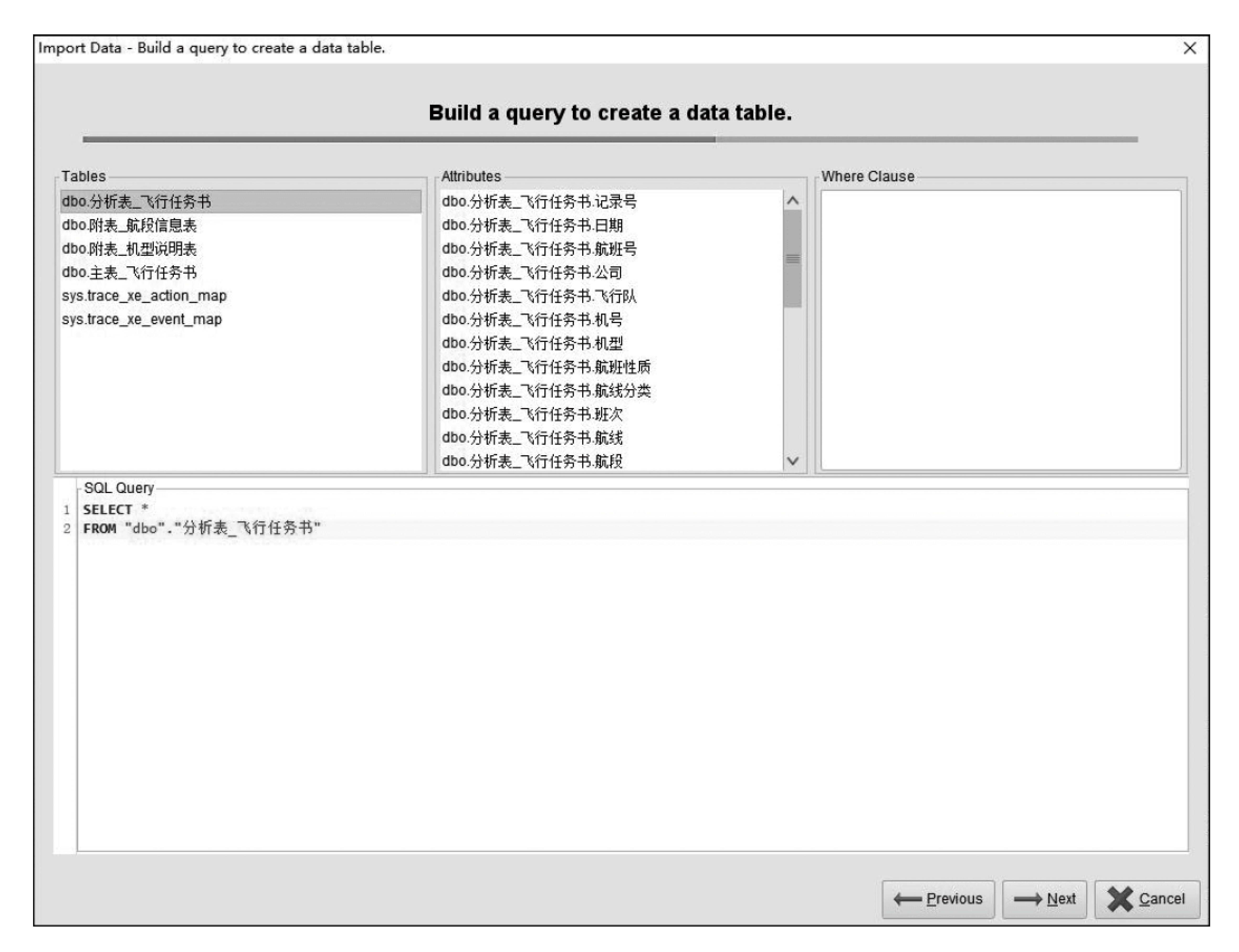


图 6-4　进一步选择需要提取的数据项

表 6-4　数据库连接参数设置

参数名称	中文释义	本案例的输入值
Name	与数据库连接的名称	Audit data
Database system	所连接的数据库类型	Microsoft SQL Server(JTDS)
Host	数据库的 IP 地址	Localhost
Port	数据库的连接端口	1433
Database scheme	数据库名称	某航空公司挖掘分析数据

(4) 单击图 6-4 上的“Next”按钮，软件弹出如图 6-5 所示的窗口，提示选择保存数据的位置。本案例选择保存在“Local Repository”下的“Data”文件夹中，输入数据的名字为“某航空公司挖掘分析数据”。

(5) 构建聚类分析流程。如图 6-6 所示，首先从“Repository”中把刚才保存在“Data”中的数据库拖放到流程窗口，再从分析处理模块中找到“Select Attributes”模块和“K-Means”聚类算法模块，把它们分别拖放到流程窗口中。“K-Means”聚类算法模块在流程窗口中显示的标识是“Clustering”。用连线按照如图 6-6 所示的样例把数据库模块、“Select Attributes”模块和“K-Means”聚类算法模块连接起来。注意：“K-Means”聚类算法模块和流程窗口最右侧的输出接口也要用连线连接起来。

(6) 配置参数。首先配置“Select Attributes”模块。“Select Attributes”模块的作用是从数据库中进一步选择“K-Means”聚类算法所需要的属性。在如图 6-6 所示的流程设计窗口中选择“Select Attributes”模块，显示“Select Attributes”模块参数如图 6-7 所示。

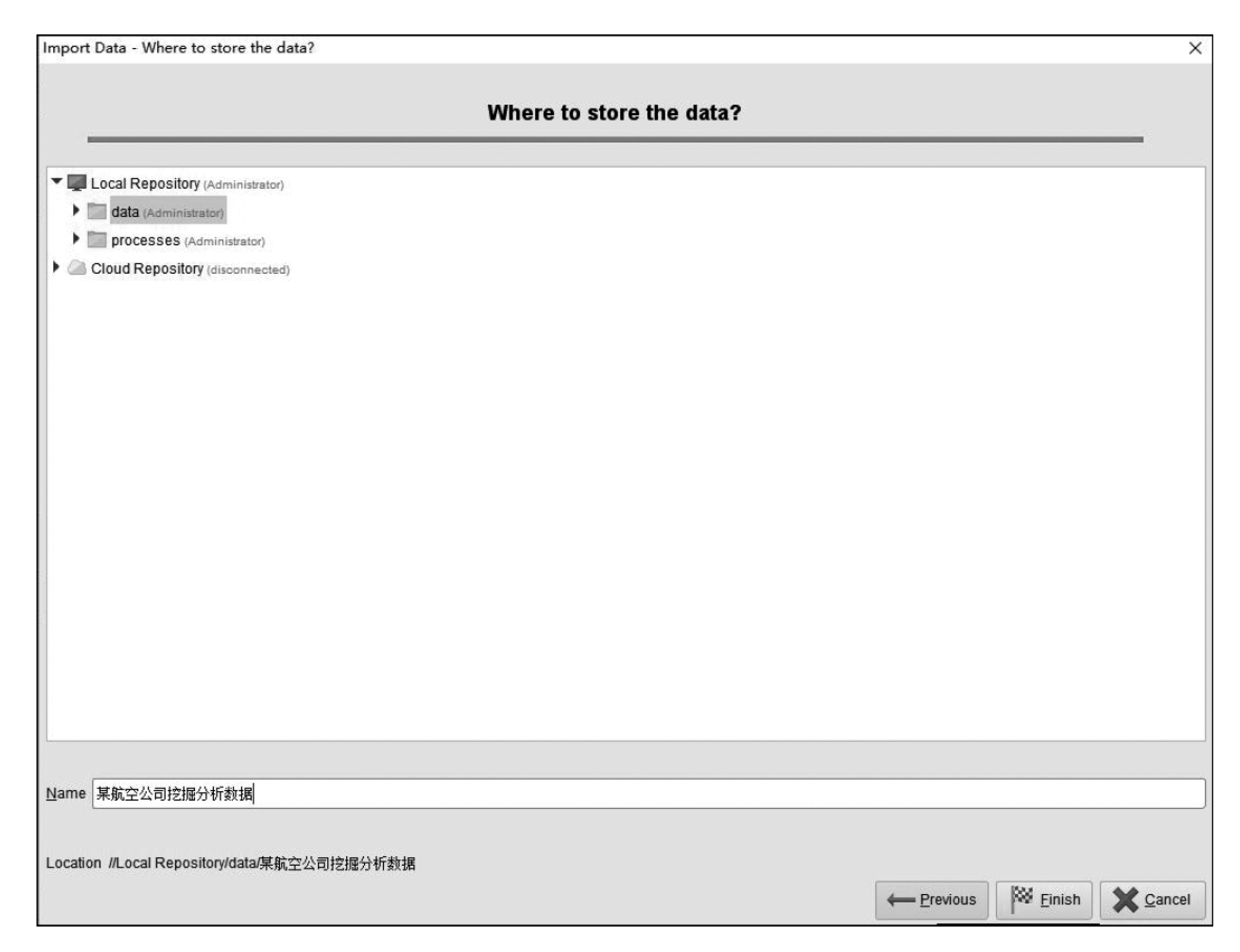


图 6-5　选择数据存放的位置

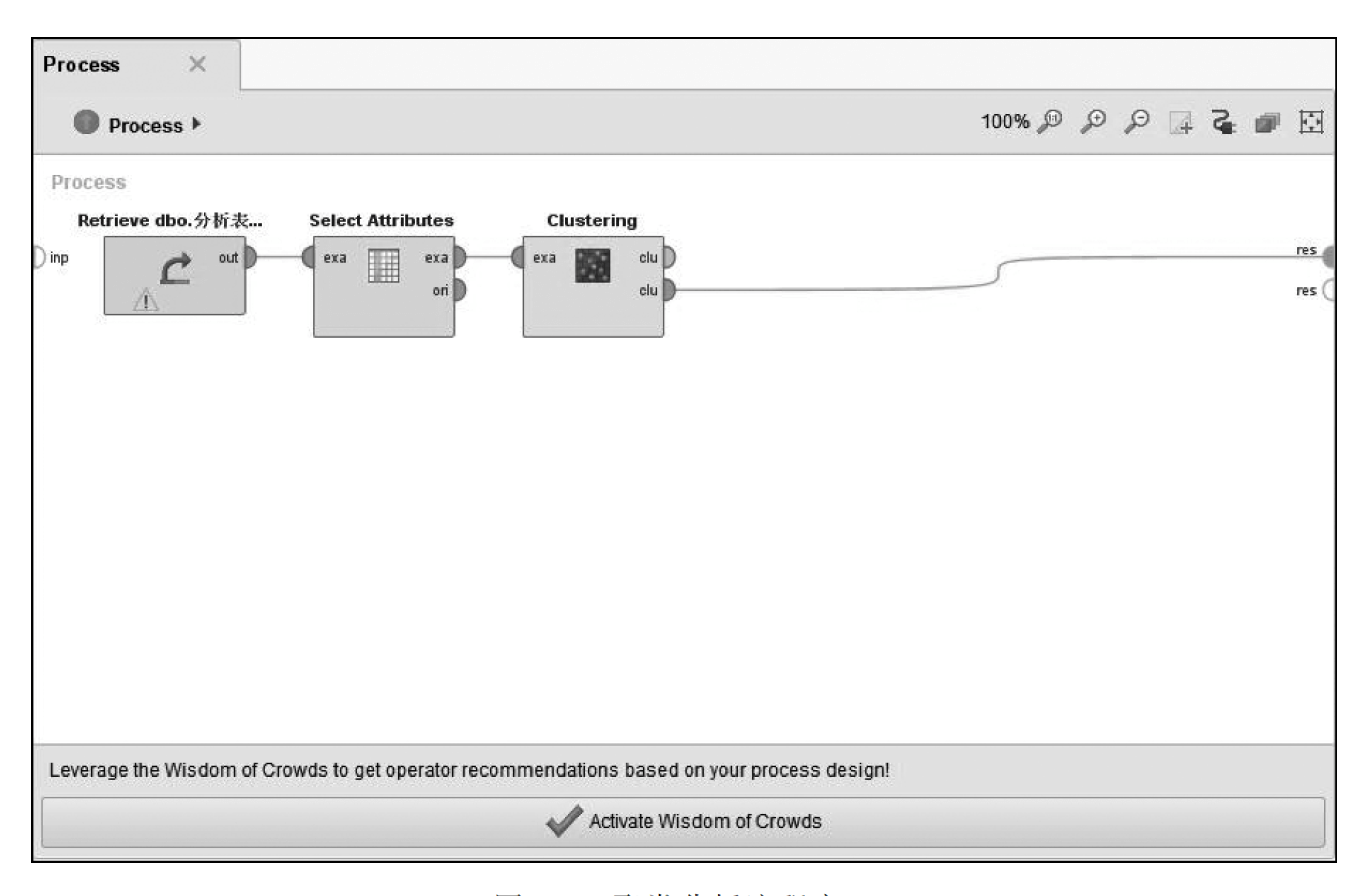


图 6-6　聚类分析流程窗口

其中主要配置的参数是“attribute filter type”，代表“选择数据属性的方式”，在本例中，我们选择“Subset”，表示要从数据库中选择一部分属性参与后续的聚类算法。当我们在“attribute filter type”参数栏选择“Subset”后，下面的“Attributes”栏自动切换成“Select Attributes”按钮，单击这个按钮，弹出如图 6-8 所示的窗口，窗口的左半部分显示数据表

的所有属性,右半部分是所选取参数与聚类算法的属性。可以通过单击图中的左右箭头添加或者移除数据属性。本案例选择"公司""飞行队""航线分类""航班性质""航段分类""航段距离"和"加耗油差额"七个属性。

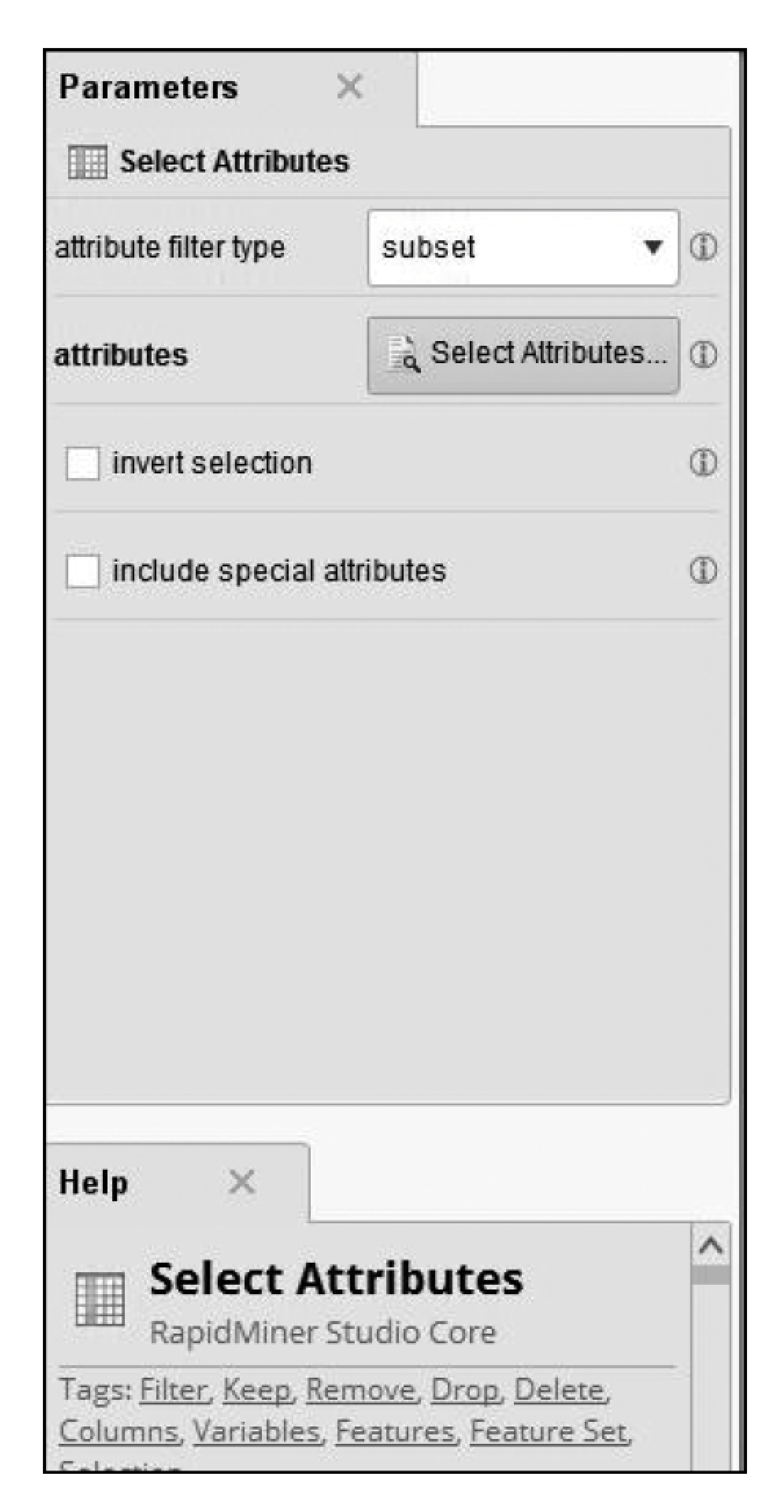


图 6-7 Select Attributes 参数配置

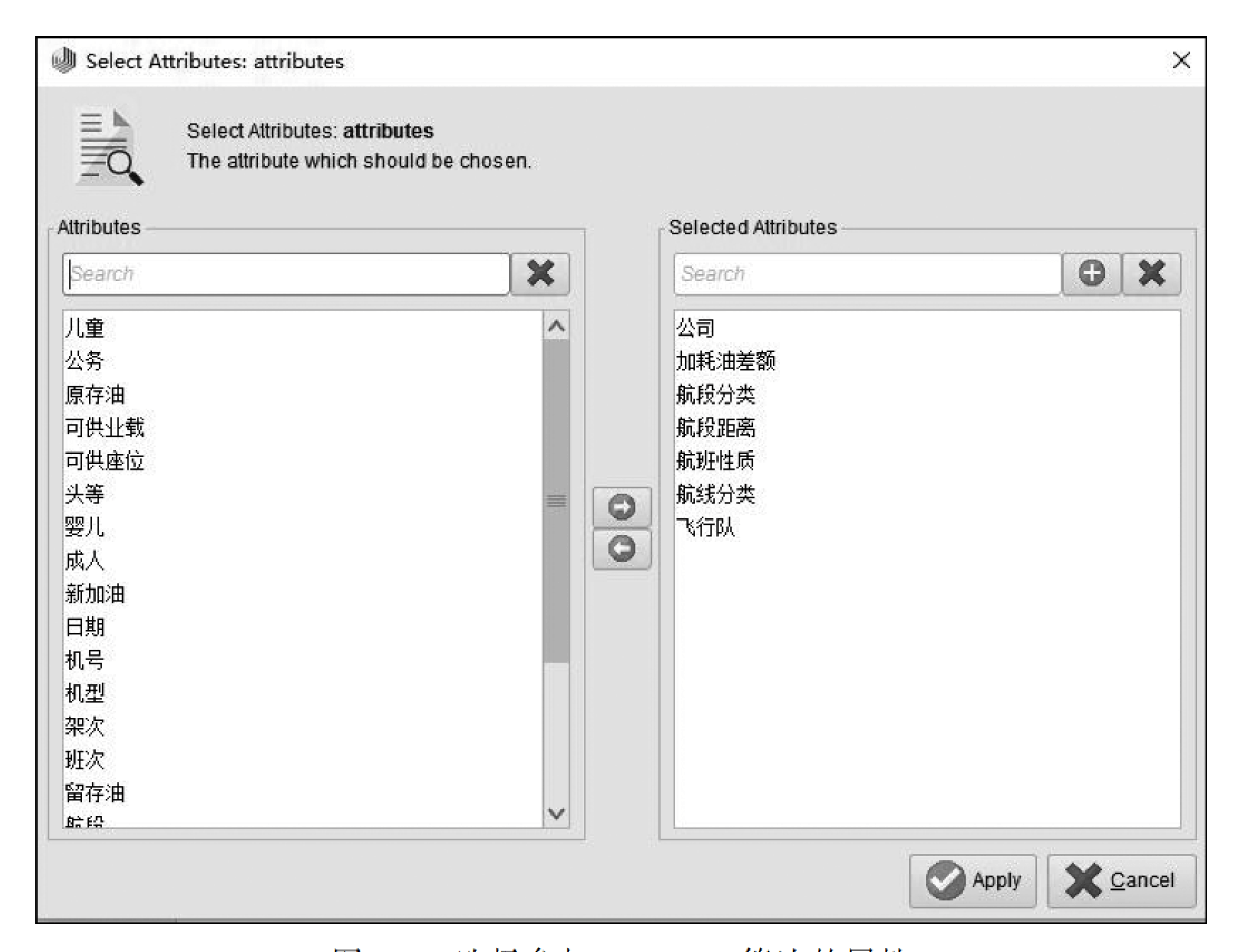


图 6-8 选择参与 K-Means 算法的属性

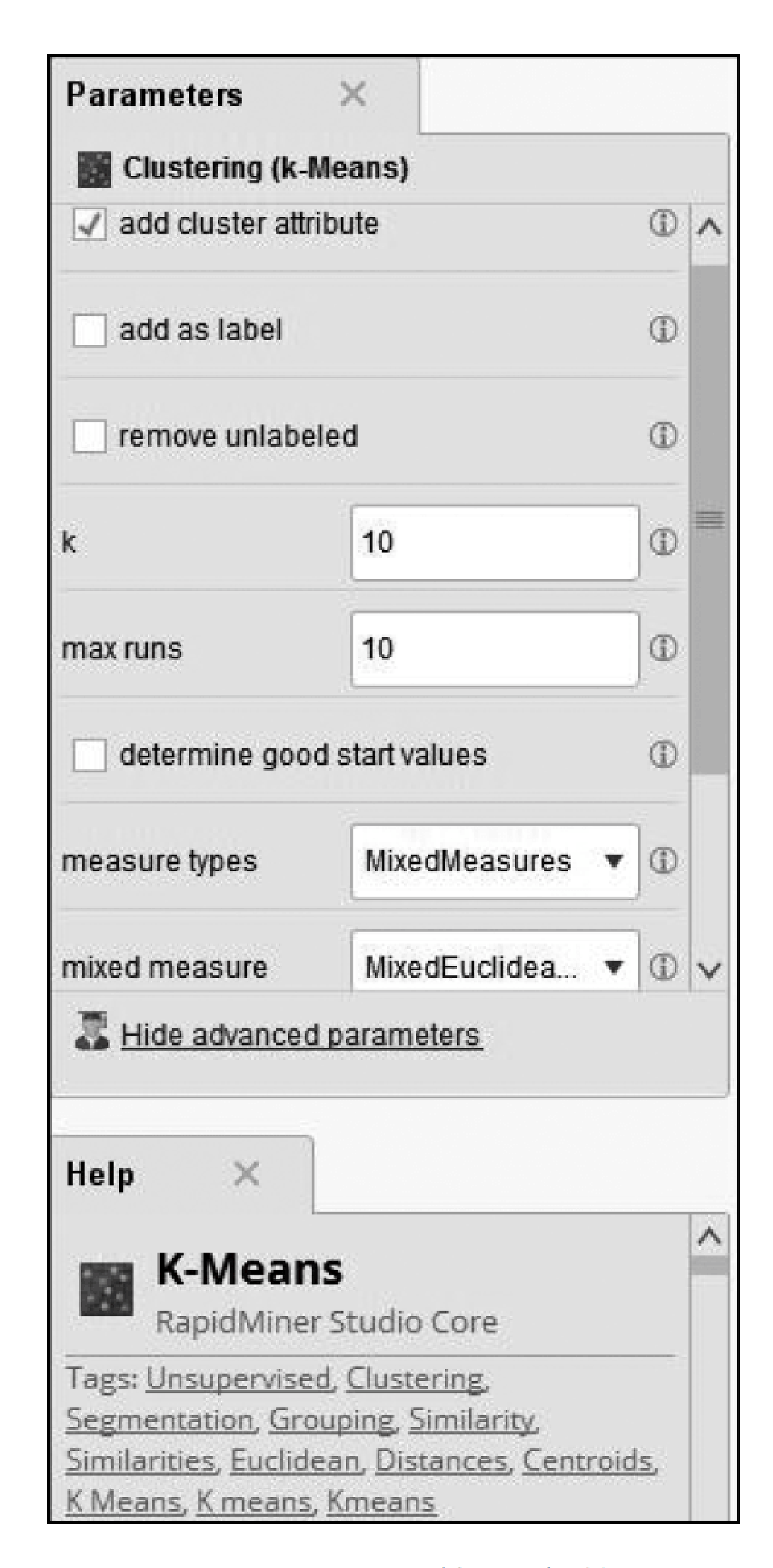


图 6-9 K-Means 算法参数配置

其次配置“K-Means”聚类算法的运行参数。在如图 6-6 所示的流程设计窗口选择“Clustering”模块,显示“Clustering”模块的参数如图 6-9 所示。其中主要的参数及其释义如表 6-5 所示。

表 6-5 “K-Means”聚类算法参数

参数名称	中文释义	本案例的输入值
add cluster attribute	如果勾选这个选项,在聚类结果中会显示“cluster”(簇)属性及其取值	选择在聚类结果中显示“cluster”(簇)属性及其取值
k	设置把数据分成几个簇	K 设置成 10,表示把数据分成 10 个簇
measure type	选择聚类的计算方式	MixedMeasures,表示选择混合欧式距离计算

(7) 单击流程的运行按钮,启动数据分析,获得聚类分析的结果,如图 6-10 所示。本案例中,聚类分析结果中出现了“cluster”一列,数据被分为 10 类(cluster0～cluster9),如图 6-10 所示的是“分类 9”的特征。从图中可以看出,这个分类中存在加油量大于消耗油的现象(加耗油差额大于 30 000),并且“航段分类”和“航线分类”这两个属性的特征值均为“国际”,说明该航空公司可能存在从境外带油的行为。

此外,RapidMiner 还对数据结果提供了丰富的统计和可视化工具,如图 6-10 最左边一栏所示。主要是“Data”“Statistics”“Charts”和“Advanced Charts”。下面简单介绍它们

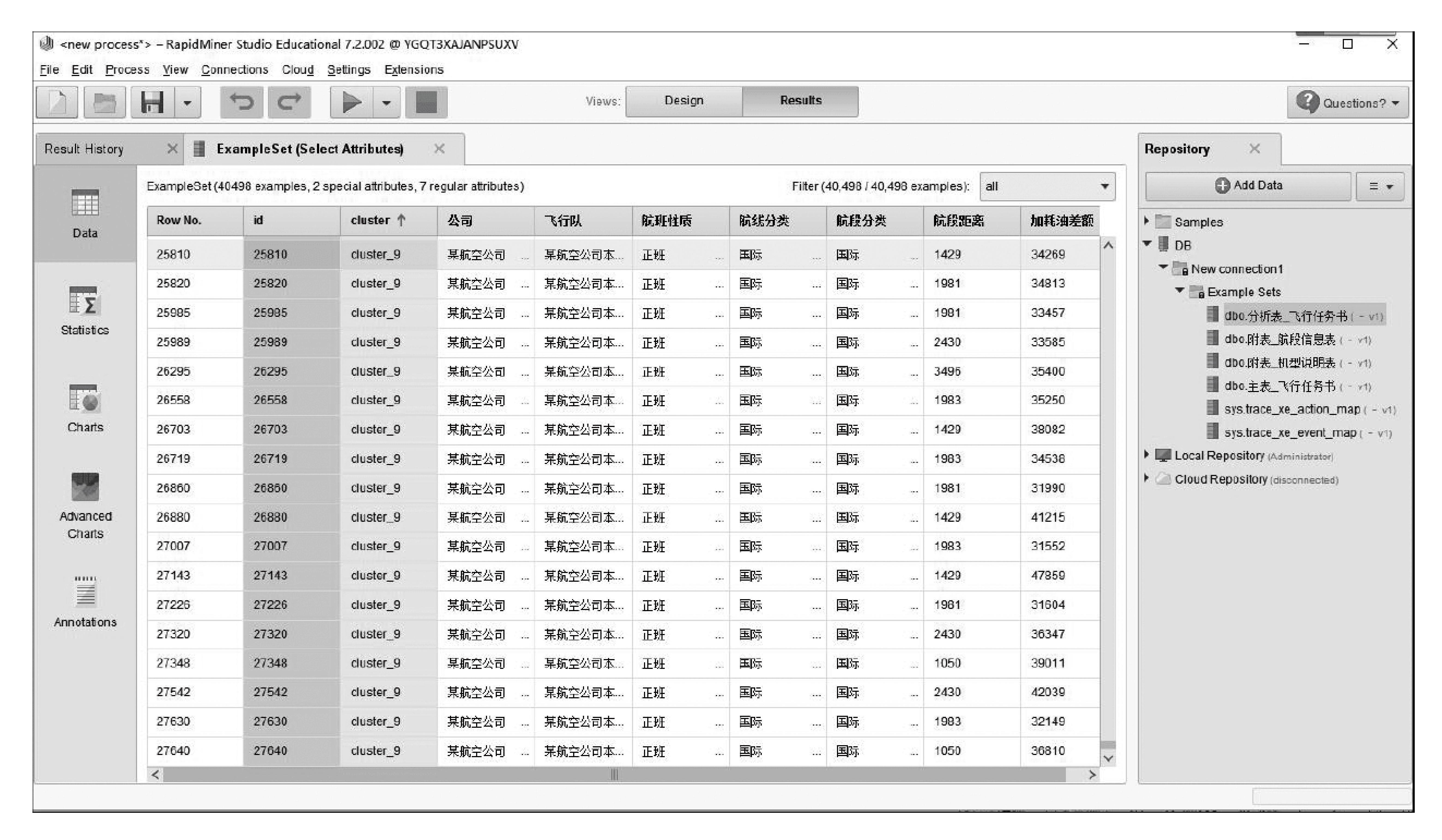


图 6-10　聚类分析的结果

的使用方法。

选择“Data”,系统将把结果以数据列表的形式显示出来,如图 6-10 所示。用户可以单击数据列表中的列名来实现排序。例如,单击“cluster”,数据列表将会按照“cluster”以升序或者降序的形式实现排序。

RapidMiner 还能够对数据结果进行统计。单击图 6-10 中的“Statistics”,弹出如图 6-11 所示的统计窗口。窗口显示出对数据表中各个属性的统计结果。单击其中的一个属性,会进一步显示详细的统计信息,例如,单击选择 “cluster”,显示关于“cluster”的统计结果,其中第 9 个簇(cluster 9)中的记录数最少,有 69 条;第 7 个簇(cluster 7)的记录数最多,有 27 637 条。再单击“cluster”中的“Values”,系统会统计并显示各个簇的记录数,如图 6-12 所示。通过这个统计表,我们可以清楚掌握各个簇中的记录数目以及它们在总数中的占比。例如,第 7 个簇(cluster 7)的记录数最多,有 27 637 条记录,占总数的 68.2%。

在图 6-10 中选择“Charts”,系统将会显示出丰富的可视化分析窗口,如图 6-13 所示。其中“Chart style”代表可以选择的统计图形种类,如图 6-14 所示。可以根据需要选择散点图、直方图、饼图、折线图等 33 种统计图形。“Plots”显示数据属性,选择其中的一个,系统将会按照“Chart style”中所选择的统计图形显示出这个属性的所有数据。例如,图 6-13 中,在“Chart style”中选择直方图(Histogram),在“Plots”中选择“cluster”属性,系统右边将显示关于各个簇中记录条数的直方图。从图中我们可以直观地看出第 7 个簇(cluster 7)中的记录数最多,而第 9 个簇(cluster 9)中的记录数最少。

图 6-10 中的“Advanced Charts”提供了更加丰富和灵活的可视化分析功能,用户可以自己定义多种维度的数据显示效果,也可以在一个图中对比显示多个属性的数据特征,如图 6-15 所示。图中“Attributes”表示数据的属性,“Chart configuration”表示如何



图 6-11　“cluster”属性的统计结果

Nominal values

Index	Nominal value	Absolute count	Fraction
1	cluster_7	27637	0.682
2	cluster_8	6034	0.149
3	cluster_6	2616	0.065
4	cluster_5	1196	0.030
5	cluster_1	1068	0.026
6	cluster_0	624	0.015
7	cluster_4	556	0.014
8	cluster_2	363	0.009
9	cluster_3	335	0.008
10	cluster_9	69	0.002

Close

图 6-12　各个簇中的记录数统计

设置图形显示参数，其中“Domain dimension”代表横坐标轴，“Numerical axis”代表纵坐标轴。我们可以从“Attributes”中选择属性拖放到“Chart configuration”中作为坐标轴的数据。例如，图中我们把属性“cluster”拖放到“Domain dimension”作为横坐标，把属性“加耗油差额”拖放到“Numerical axis”作为纵坐标，右边的坐标图将显示各个簇中加耗油差额的分布。从坐标图中可以看到，第 10 个簇(cluster 9)、第 4 个簇(cluster 3)和第 5 个簇(cluster 4)中“加耗油差额”存在大于 15 000 的情况，可能存在从境外带油的

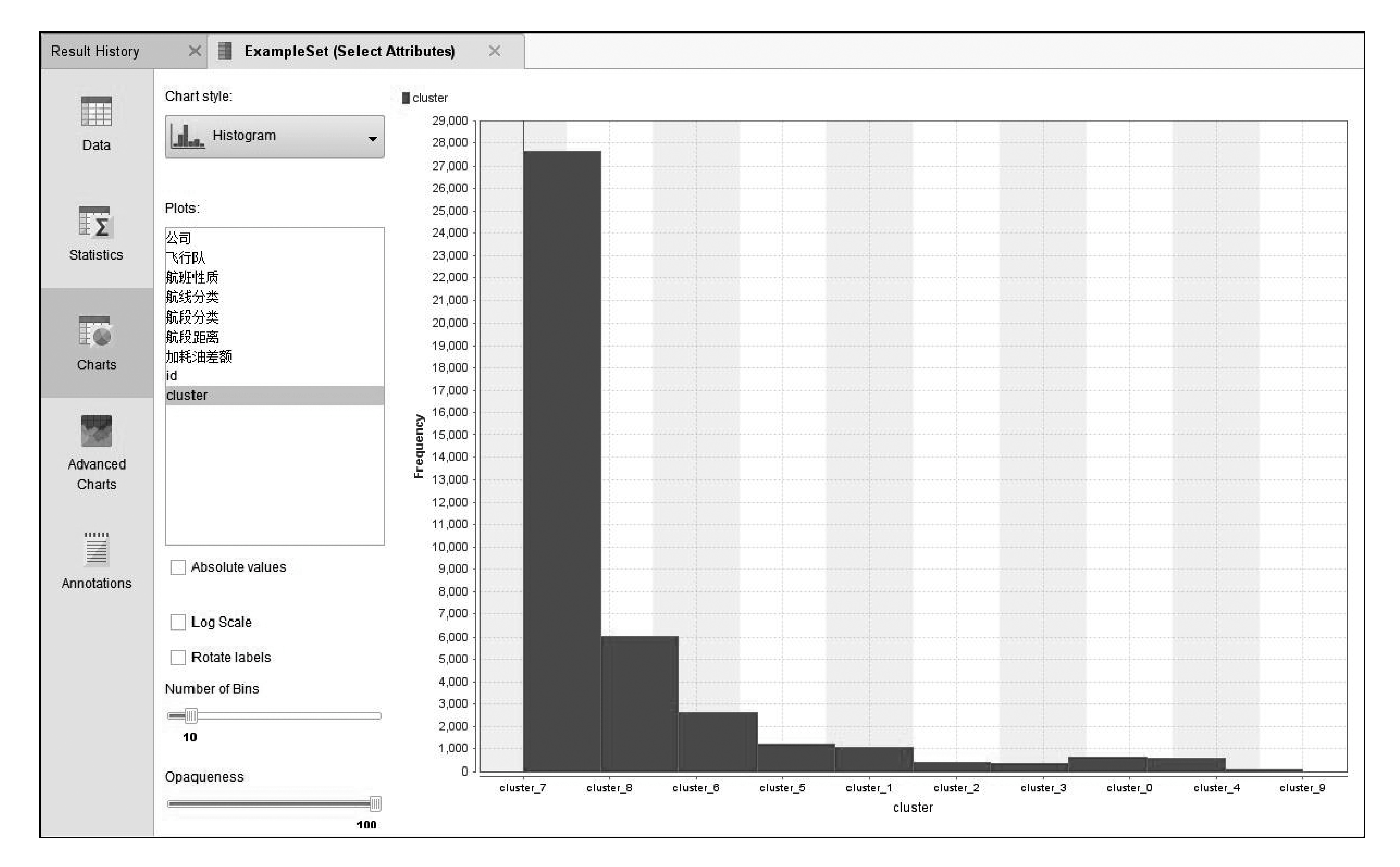


图 6-13　可视化分析示例

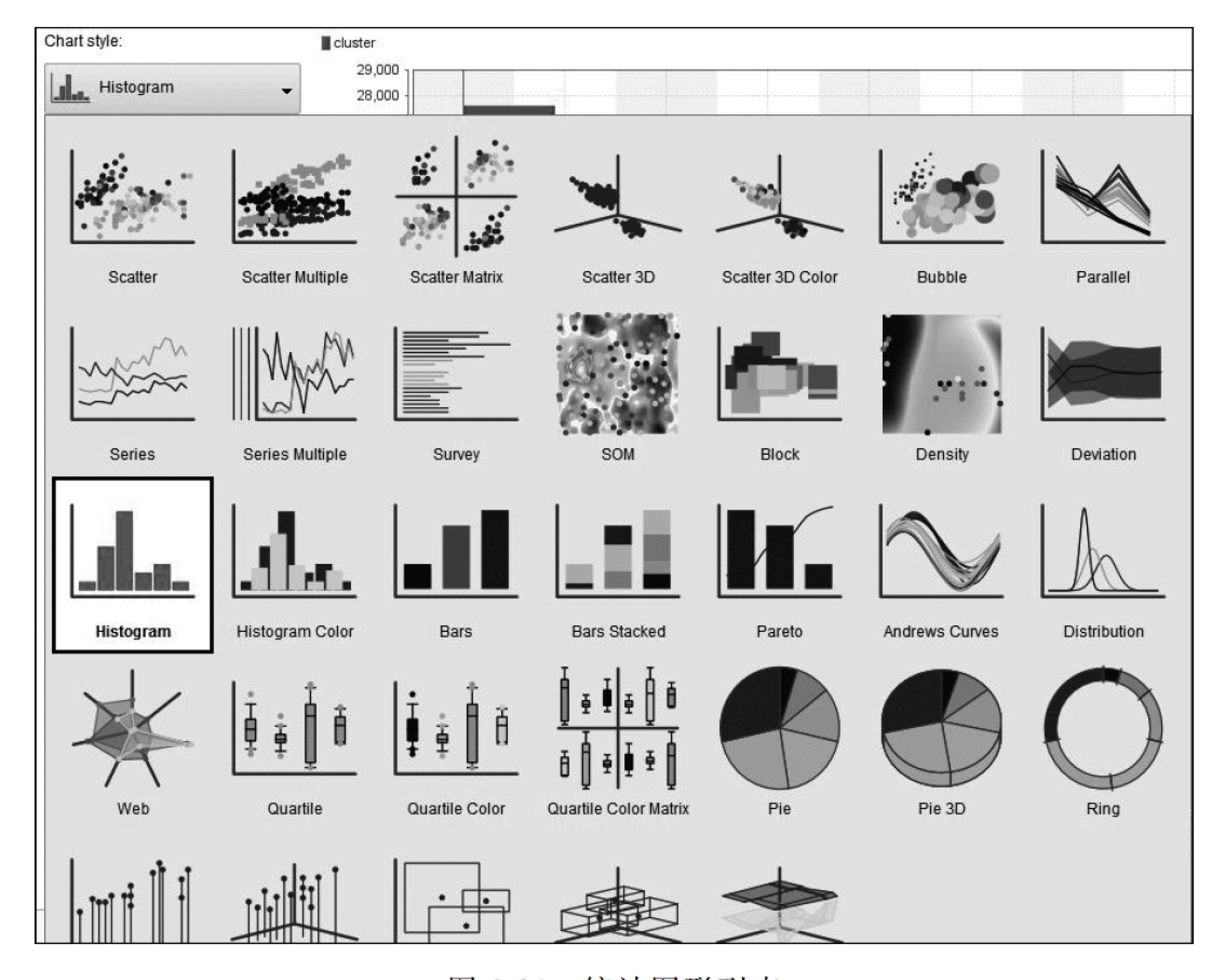


图 6-14　统计图形列表

行为。因此审计人员可以重点对这三个簇中的数据进行分析查证。通过这个例子，我们可以看到，借助 RapidMiner 强大的可视化的手段，审计人员可以很容易地锁定审计分析的重点数据。

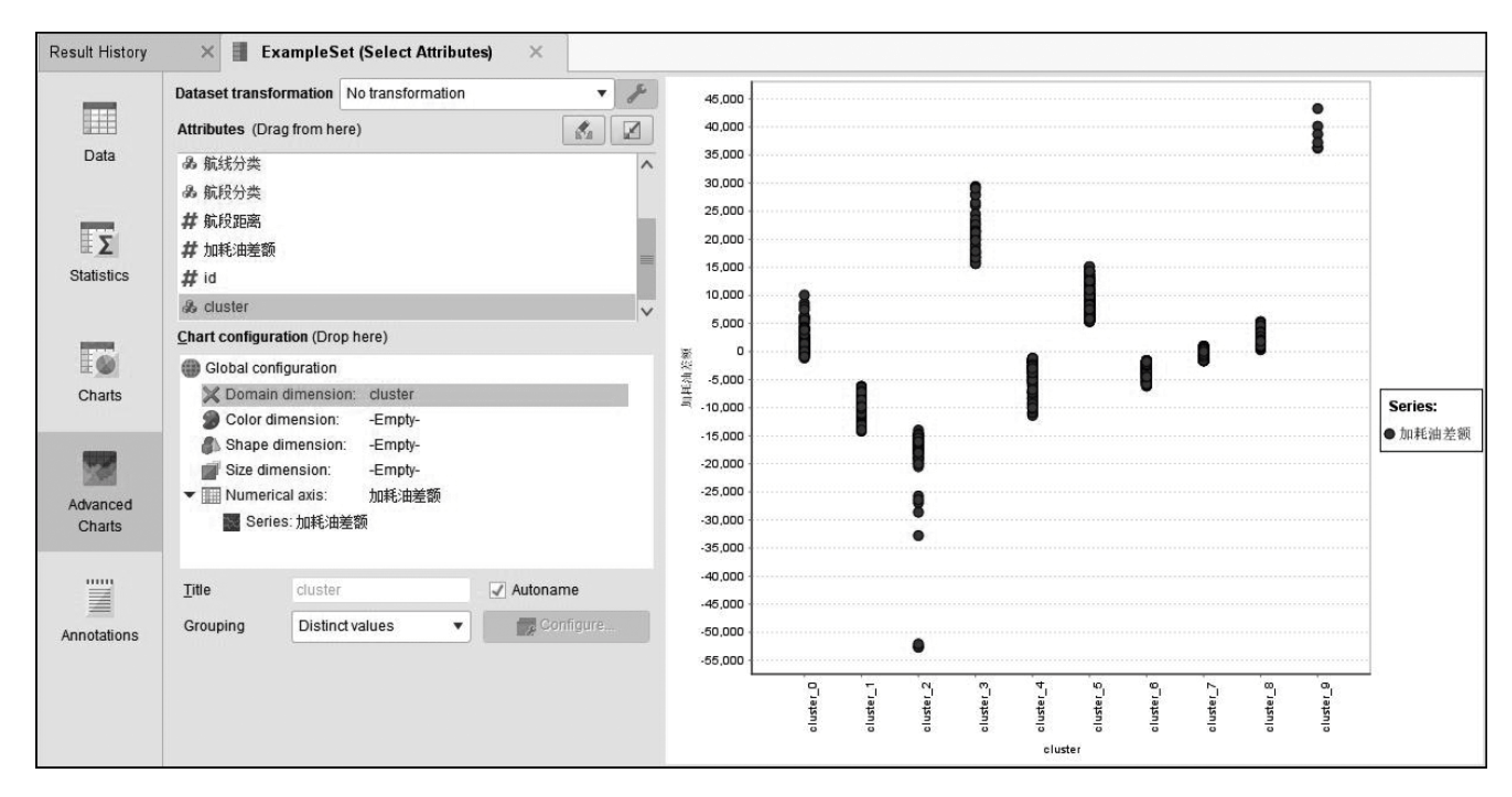


图 6-15　高级的可视化分析窗口

6.2　挖掘分析在推荐系统中的应用

6.2.1　案例背景

随着电子商务规模的不断扩大，商品数量和种类快速增长，顾客需要花费大量的时间才能找到自己心仪的商品。为了解决这个问题，推荐系统应运而生。推荐系统是建立在海量数据挖掘基础上的一种商务智能平台，它是根据顾客的需求、兴趣等，将顾客感兴趣的商品推荐给顾客的个性化服务系统。

基于用户相似度的协同过滤推荐系统是目前广泛使用的一种推荐系统，它推荐商品的原理是“跟你喜好相似的人喜欢的东西你也很有可能喜欢”。构建这种推荐系统需要分成两个阶段，首先是建立推荐模型，利用顾客已知的消费数据和推荐算法，通过训练获得推荐模型；其次是推荐阶段，使用所构建的推荐模型对其他顾客进行相关商品的推荐。

现有一商家在其电子商务网站拥有顾客对各种商品的评分数据，利用 RapidMiner 帮助该商家构建基于用户相似度的协同过滤推荐系统，以便向其他爱好相似的顾客推荐相关商品。

6.2.2　数据准备

首先需要采集数据，利用网络爬虫从电子商务网站上获取顾客对商品的评分数据，然后对这些数据进行清洗和整理后，我们获得两个数据集：一个是用于构建推荐模型的建模数据集；另一个是测试数据集。利用构建好的推荐模型对测试数据集中的顾客推荐相关的商品。这两个数据集都只包含顾客编号、数据编号及商品评分三个属性，如表 6-6 所示。

表 6-6　推荐系统使用的数据集结构

名　　称	数 据 类 型	备　　注
user_id	Int	顾客编号
item_id	Int	商品编号
score	Int	商品评分

6.2.3　构建推荐系统

下面介绍基于上述建模数据集和测试数据集构建推荐系统并进行测试的过程。

1. 导入建模数据和测试数据

(1) 启动 RapidMiner 软件,单击“Repository”标签下方的“Add Data”按钮,弹出加载数据库的窗口,如图 6-16 所示。



图 6-16　添加建模数据

(2) 在如图 6-16 所示的图中选择“My Computer”选项,软件提示输入本地计算机中建模数据所在的位置,如图 6-17 所示。找到并选择“建模数据. txt”(评分数据保存在此文件中),单击“Next”按钮,弹出如图 6-18 所示的窗口。

(3) 在图 6-18 中单击“Next”按钮,弹出如图 6-19 所示的窗口,系统给导入的建模数据分配的默认属性名分别是“att1”“att2”和“att3”,可以单击每个属性旁边的齿轮符号来修改默认的属性名和数据类型。本案例中,我们把三个默认的数据属性分别更改为“user_id”“item_id” 和“score”,分别代表顾客编号、商品编号和商品评分,如图 6-20 所示。

(4) 选择存放建模数据的位置。本案例中,导入的建模数据存放在“Local Repository”下的“Data”文件夹中,数据的名字为“建模数据”,如图 6-21 所示。单击

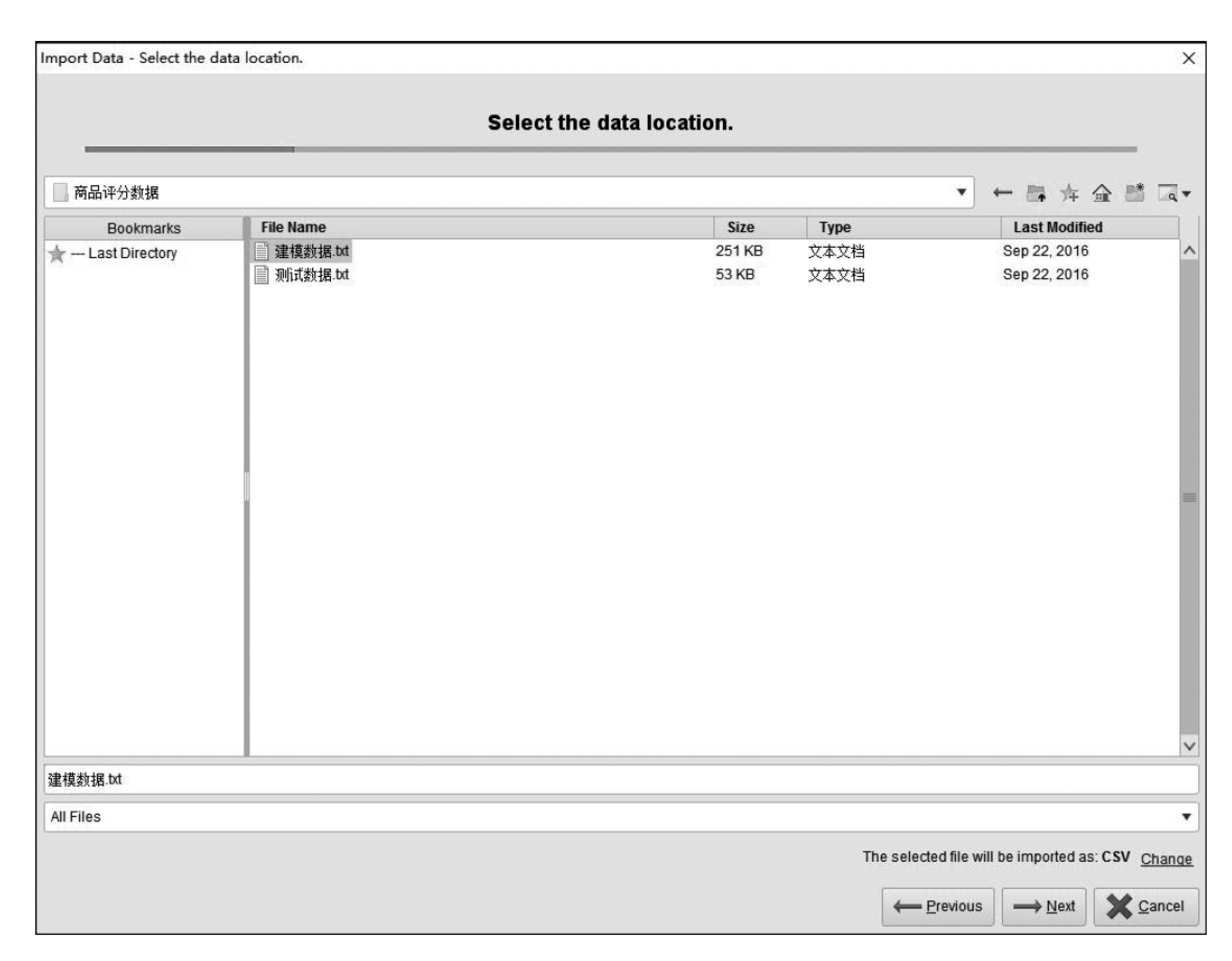


图 6-17　选择建模数据所存放的位置

Import Data - Specify your data format

Specify your data format

Header Row　1　File Encoding　GBK　Use Quotes　"

Start Row　1　Escape Character　\　Trim Lines

Column Separator　Semicolon ";"　Decimal Character　.　Skip Comments　#

1	45874270	2348372	4
2	45874270	3216007	5
3	45874270	1261560	5
4	45874270	3138847	5
5	45874270	1044177	5
6	45874270	3142118	5
7	45874270	3234345	5
8	45874270	3151575	5
9	45874270	4219500	5
10	45874270	1116367	5
11	45874270	1054889	5
12	45874270	1048173	5
13	45874270	3225658	5
14	45874270	3343988	5
15	45874270	3574119	5
16	45874270	1322025	5
17	45874270	1865089	5

PREVIEW

no problems.

Previous　Next　Cancel

图 6-18　设置数据的格式

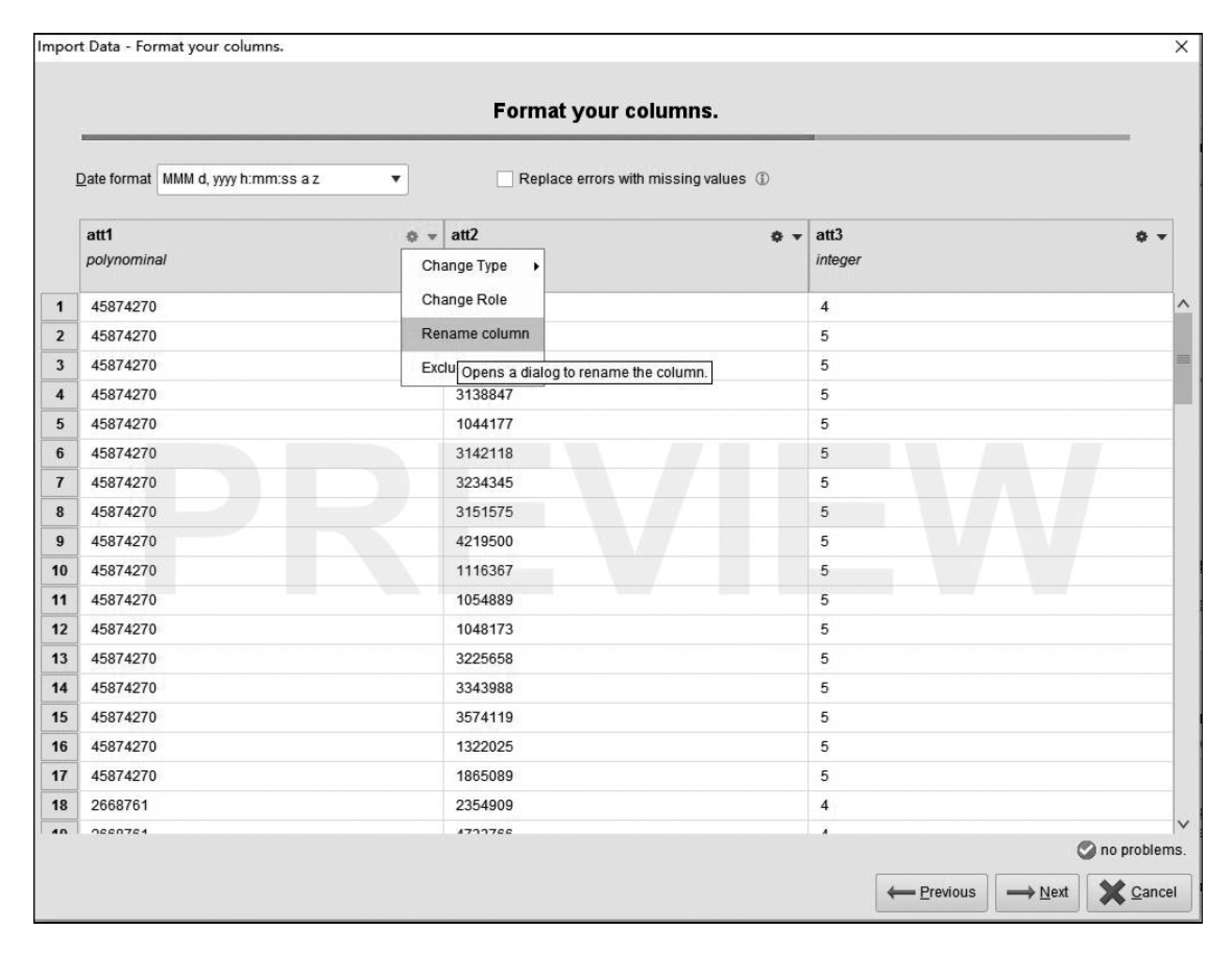


图 6-19　设置建模数据的属性名和数据类型

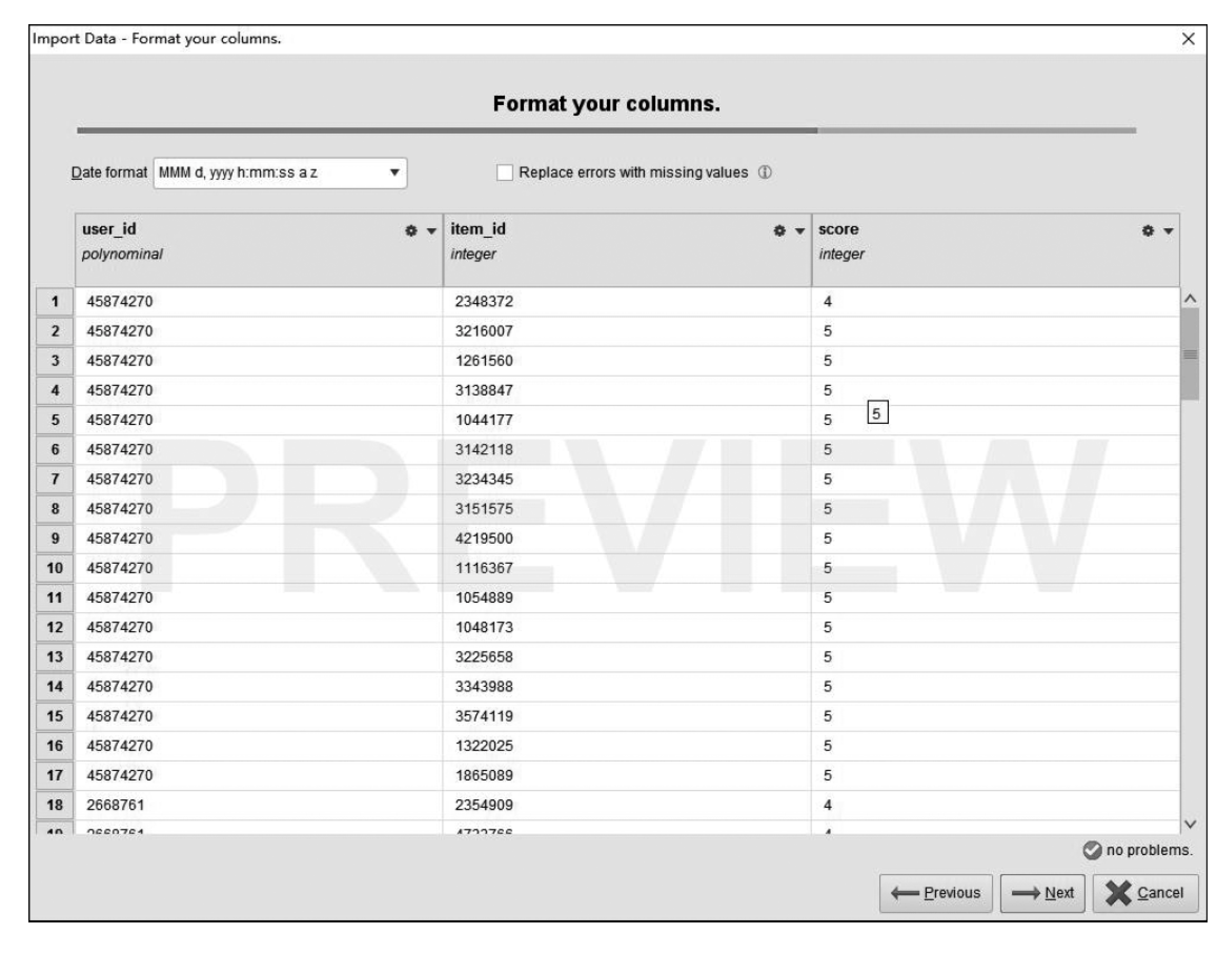


图 6-20　本案例中建模数据的属性名称

“Finish”按钮，完成导入“建模数据”。

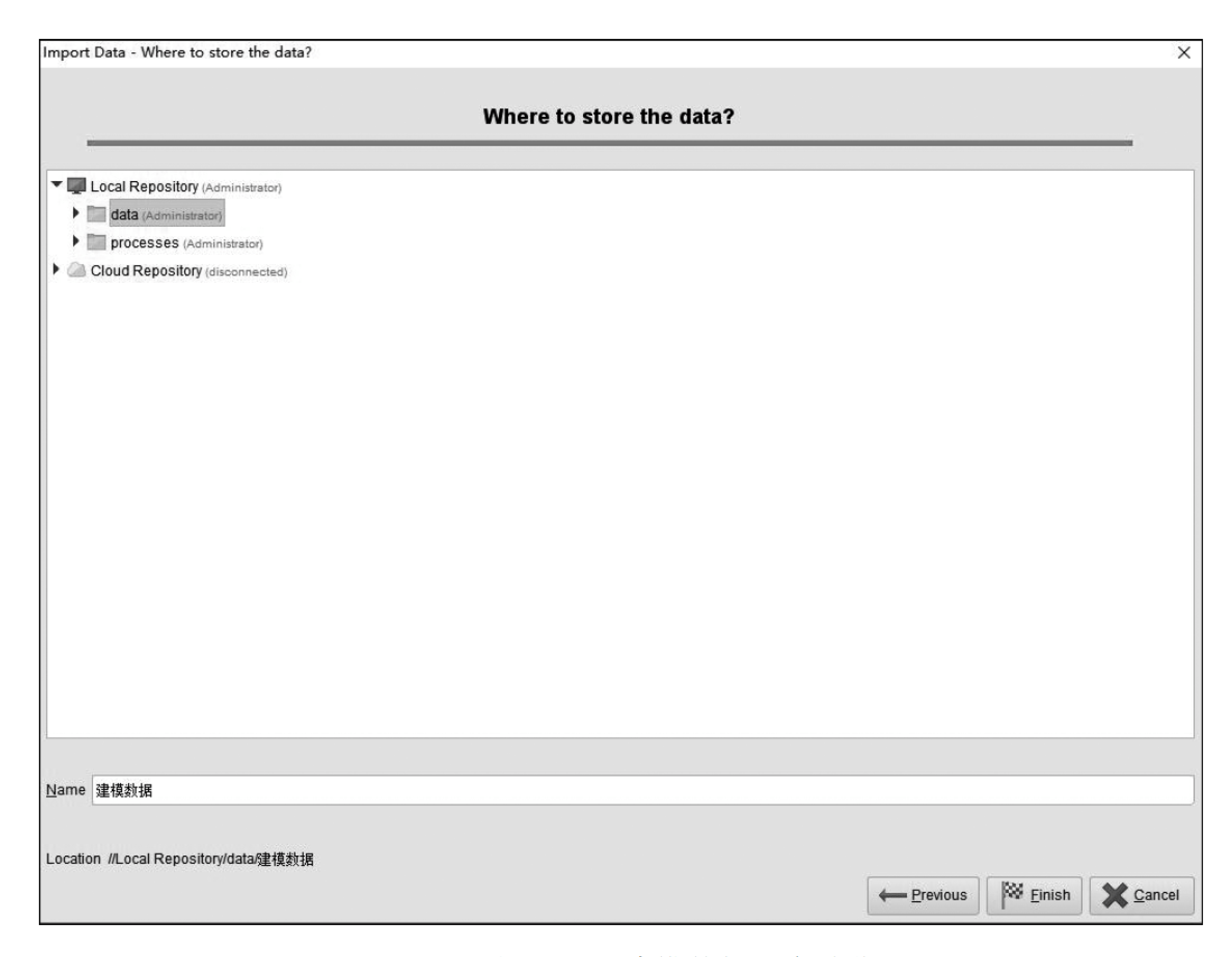


图 6-21　选择导入的建模数据的存放位置

(5) 用同样的方法导入“测试数据”后，在“Local Repository”区域会出现导入的建模数据和测试数据，如图 6-22 所示。

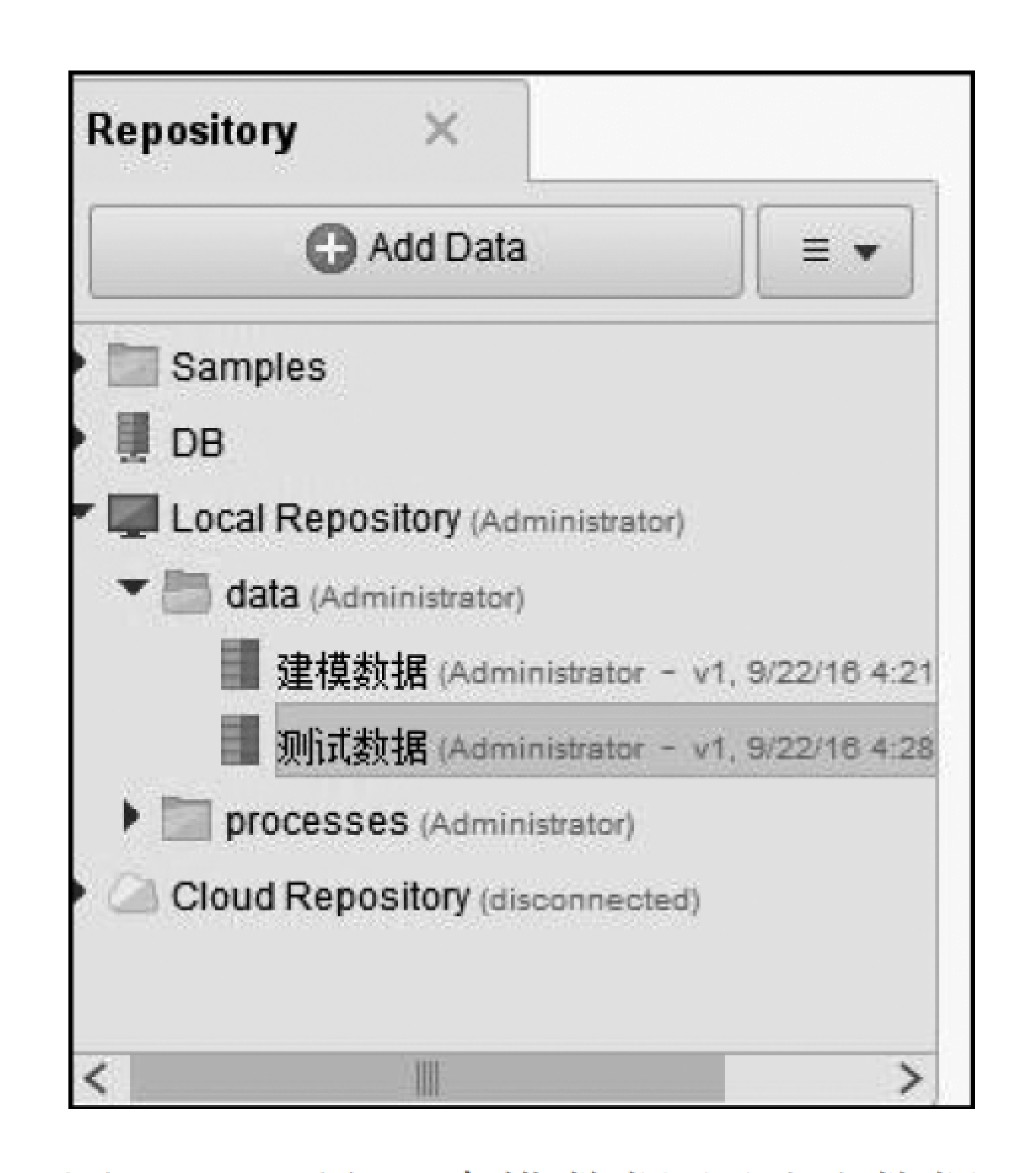


图 6-22　导入建模数据和测试数据

2. 下载并安装推荐算法模块

(1) 如图 6-23 所示，在 RapidMiner 主界面的菜单栏单击“Extensions”，在弹出的子菜单中进一步选择“MarketPlace(Updates and Extensions)”，弹出如图 6-24 所示的窗口。

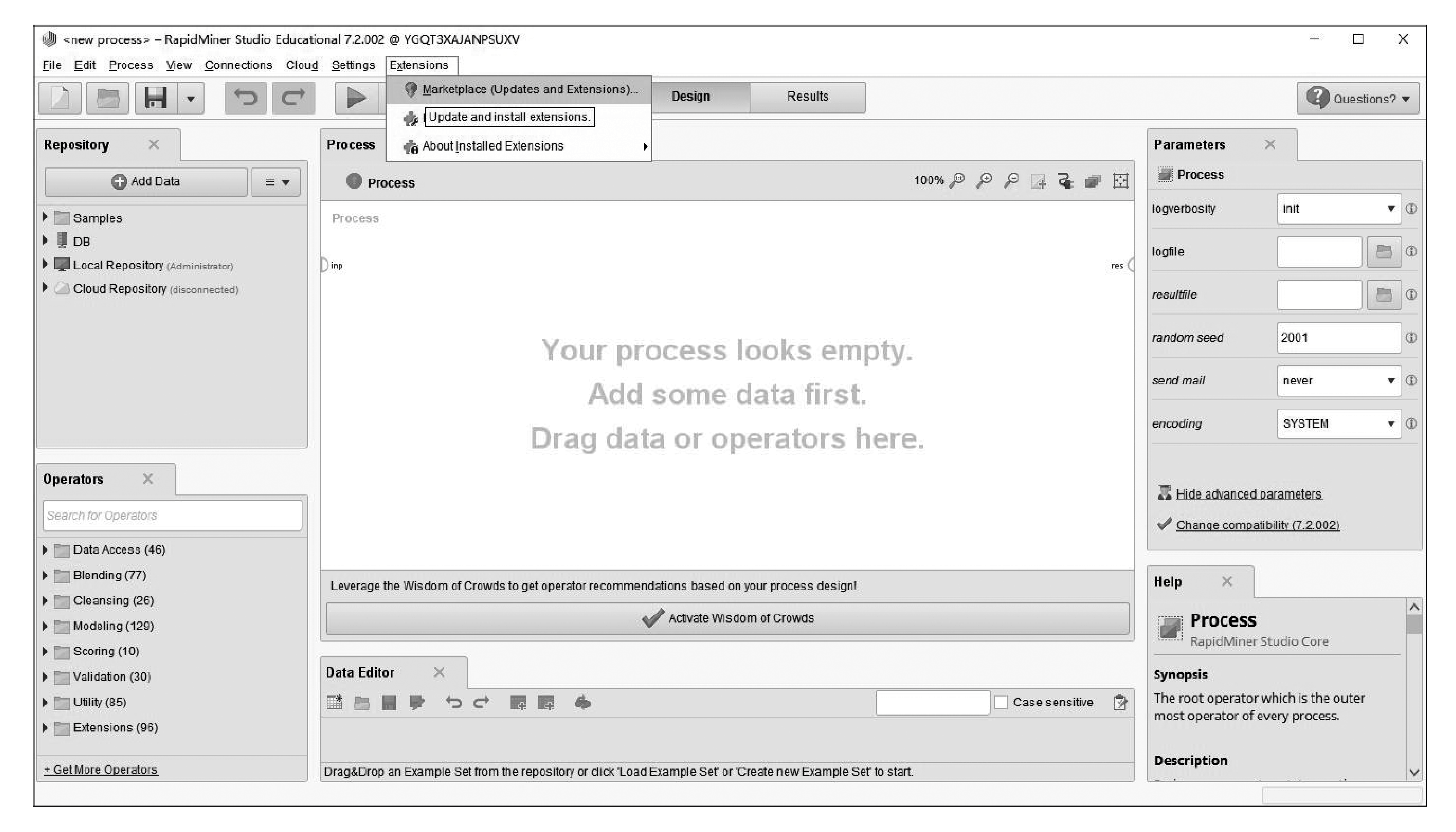


图 6-23　在 RapidMiner 主界面的菜单栏选择安装扩展模块

（2）在如图 6-24 所示的窗口中选择“Search”标签并输入“recomm”查找到“Recommender Extension5.1.2”，单击“Install Packages”按钮开始安装推荐算法模块。安装完成后，在 RapidMiner 的主界面的数据处理模块区域会出现“Recommender”，如图 6-25 所示。在后续构建推荐系统时，我们会从中选择相应的推荐算法。

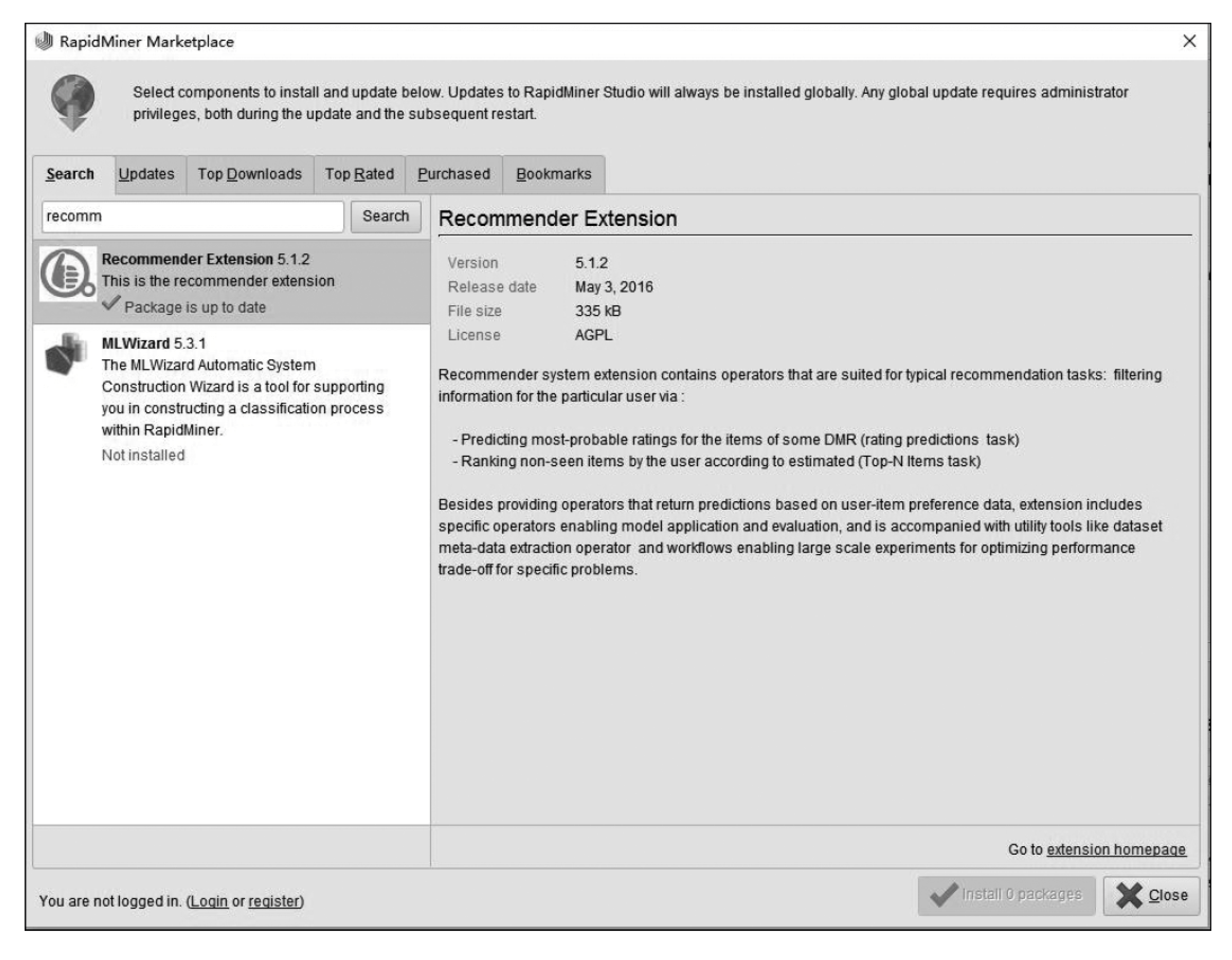


图 6-24　查找并安装推荐算法模块

3. 构建推荐系统

如图 6-26 所示，推荐系统由两部分组成：一个是建立推荐模型，选择“建模数据模块”“设置角色模块”和“User—KNN”推荐算法模块并拖放到流程窗口的上面一行，这一行的模块用于构建推荐模型；另一个是应用推荐模型来完成推荐任务，选择“测试数据模块”“设置角色模块”和“Apply Model”模块并拖放到流程窗口的下面一行中，这一行的模块利用上述构建好的推荐模型向测试数据中的顾客推荐相关的商品，其中“Apply Model”模块的作用是使用构建好的推荐模型。把各个模块按照图 6-26 中的样式用连线连接起来。

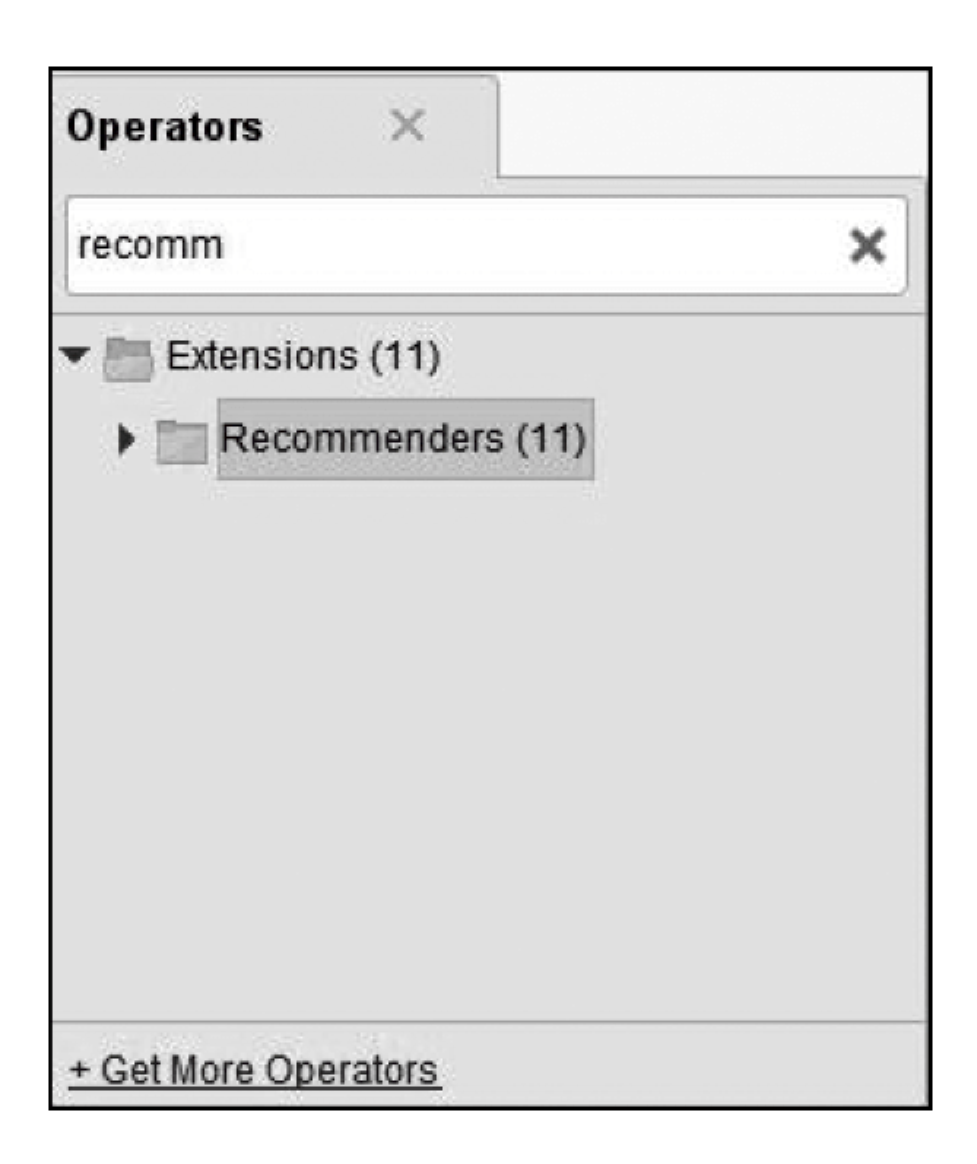


图 6-25　安装好的推荐算法模块

4. 配置参数

对图 6-26 中的各个“设置角色”模块和 “User—KNN” 推荐算法模块分别设置参数，如图 6-27～图 6-30 所示。其中图 6-27 和图 6-28 分别对构建推荐模型阶段的两个“设置角色”模块中的顾客编号和商品编号进行配置，图 6-29 是对测试阶段“设置角色”模块中的顾客编号进行配置，图 6-30 是设置“User—KNN”算法对每个顾客推荐商品的数量，本例设置为 $n=5$，意味着推荐模型只向测试数据中的每个顾客最多推荐 5 件商品。

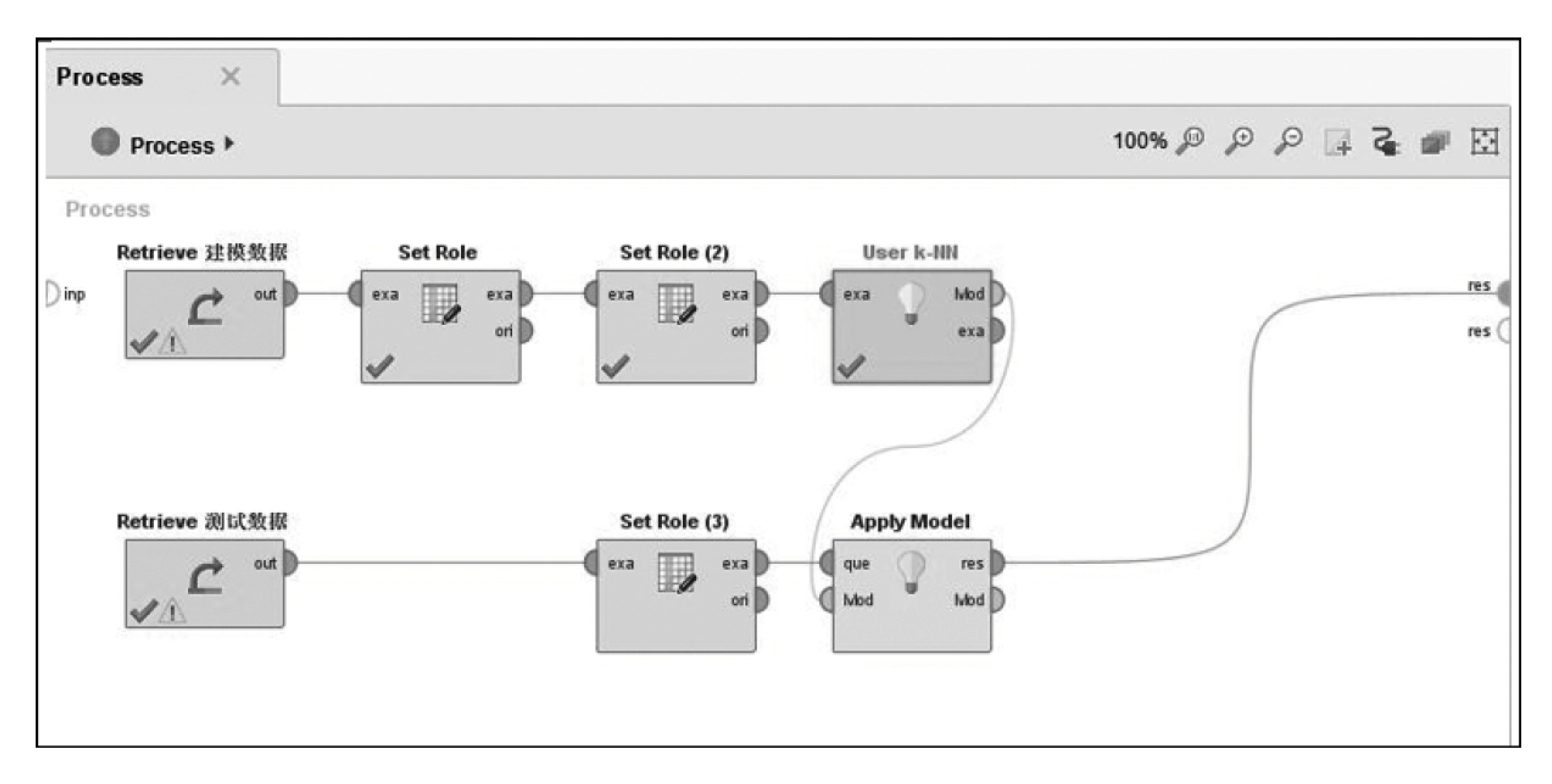


图 6-26　构建推荐系统

5. 运行流程

单击流程的运行按钮，推荐系统对测试数据集中的顾客推荐商品，推荐结果如图 6-31 所示。图中“user_id”代表顾客编号，“item_id”代表商品编号，“rank”代表推荐的优先级(1 代表最高优先级，5 代表最低优先级)。例如，图中对编号为“44240613”的顾客，推荐编号是“3323633”“1044177”“3138847”“1261560”和“4273386”5 件商品，其中“3323633”推荐的优先级最高，“4273386”推荐的优先级最低。

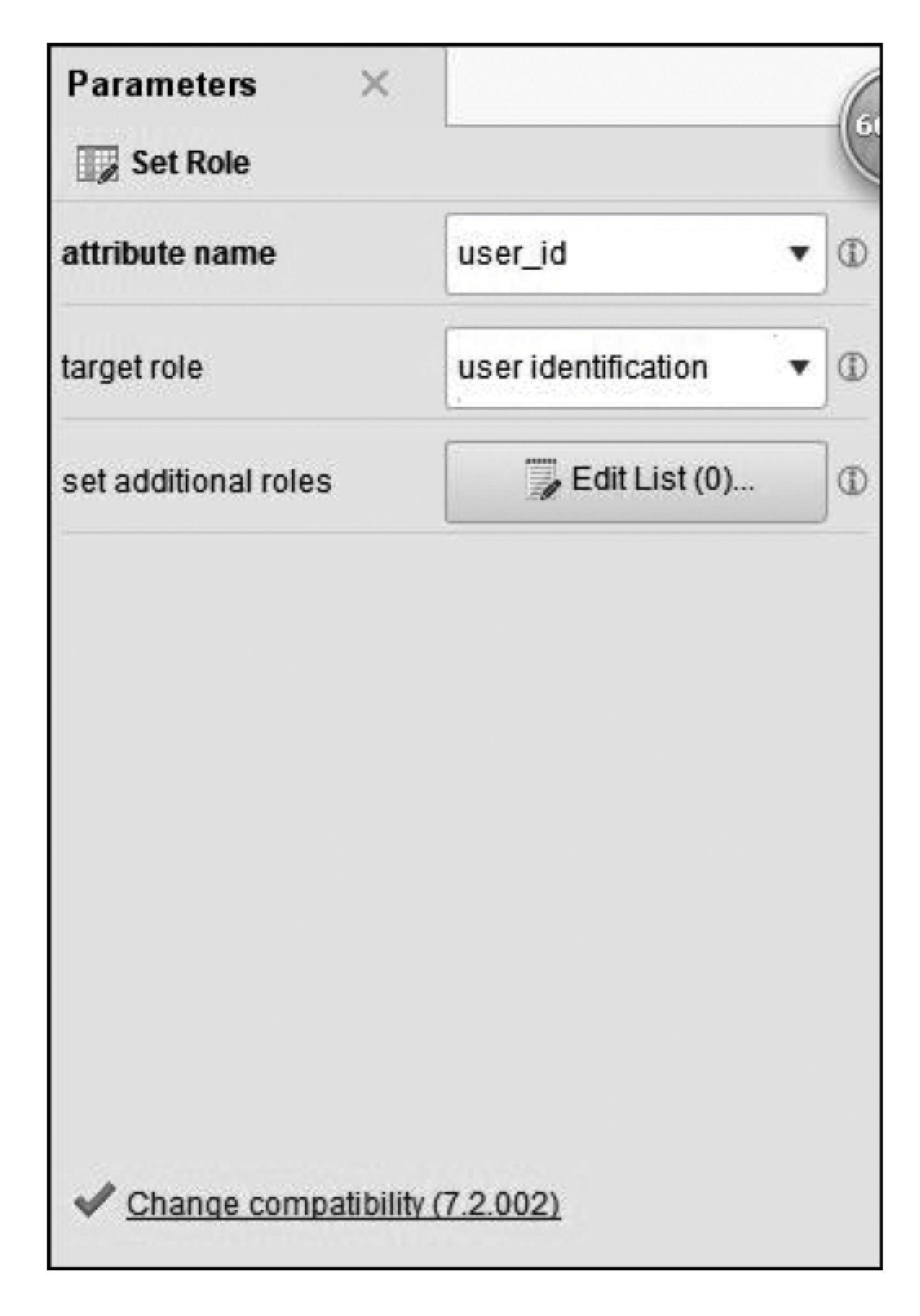


图 6-27　设置顾客编号

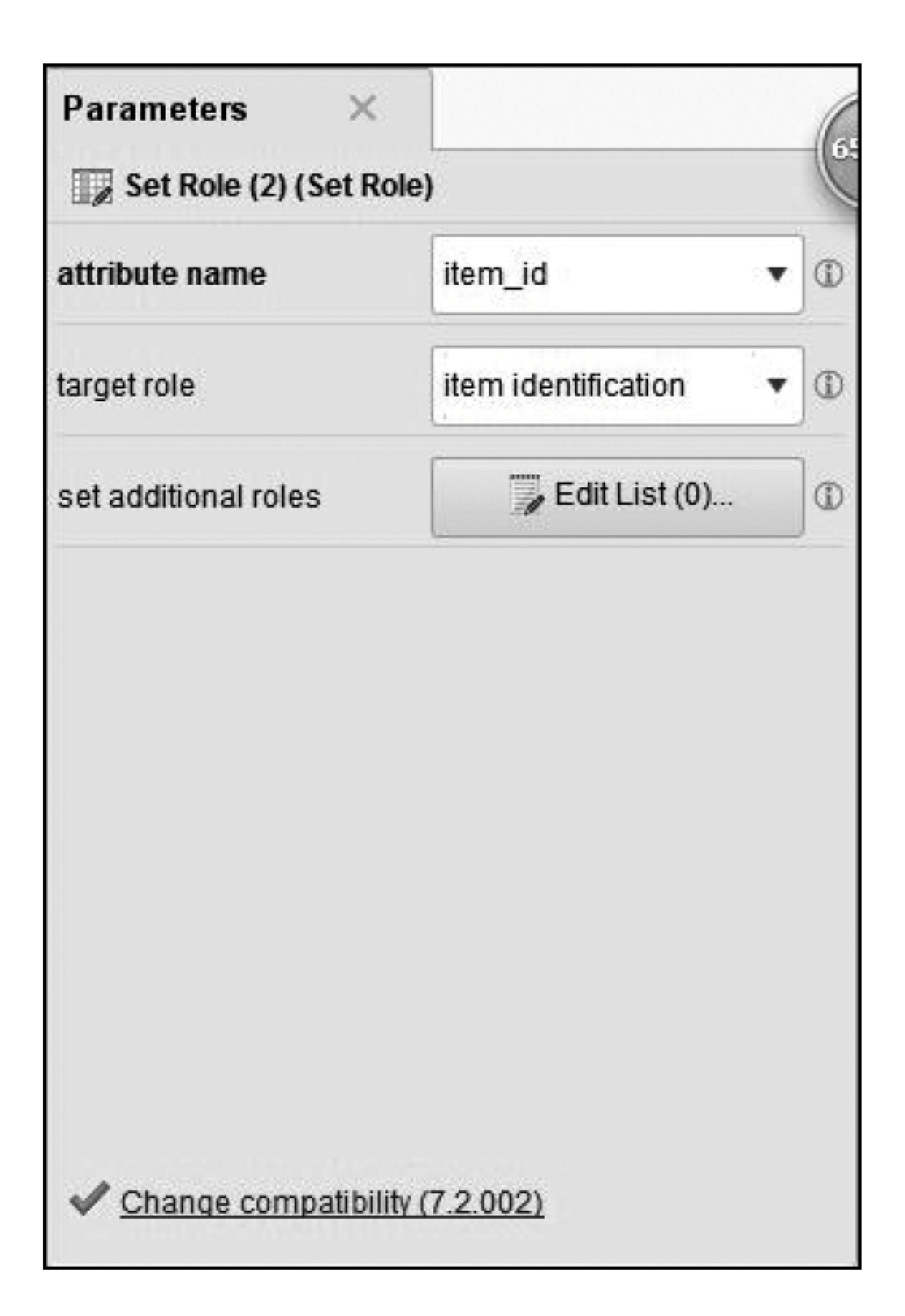


图 6-28　设置商品编号

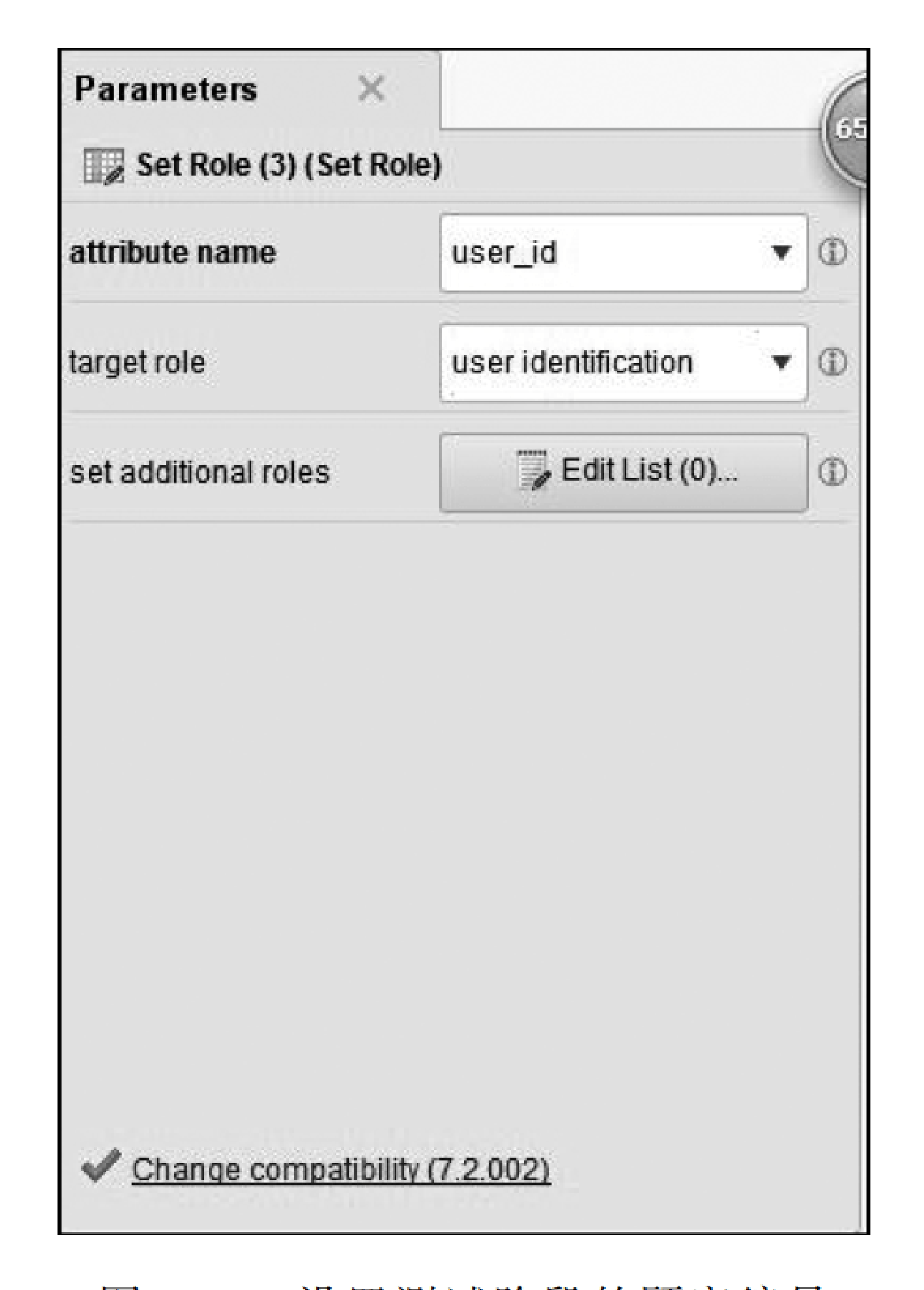


图 6-29　设置测试阶段的顾客编号

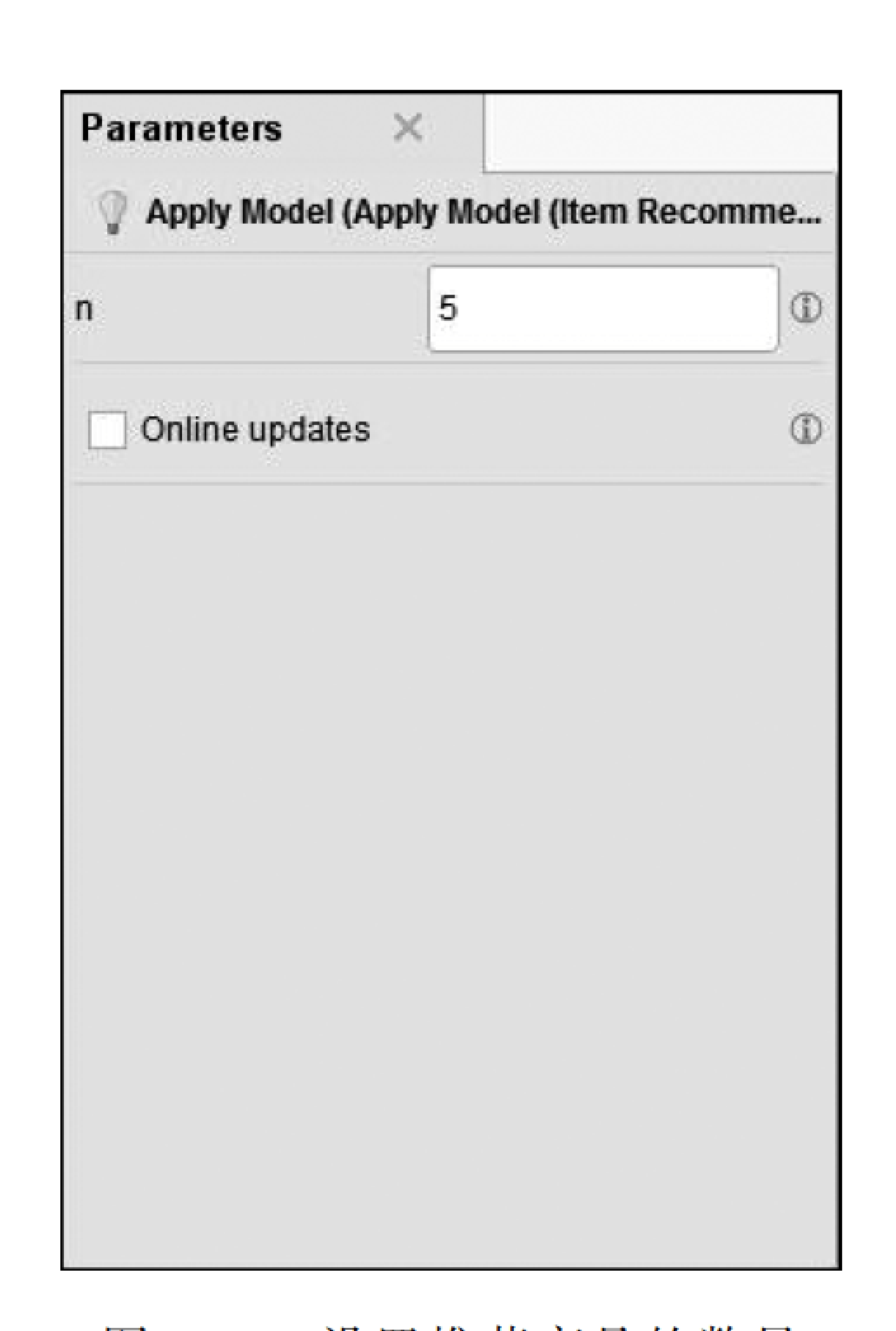
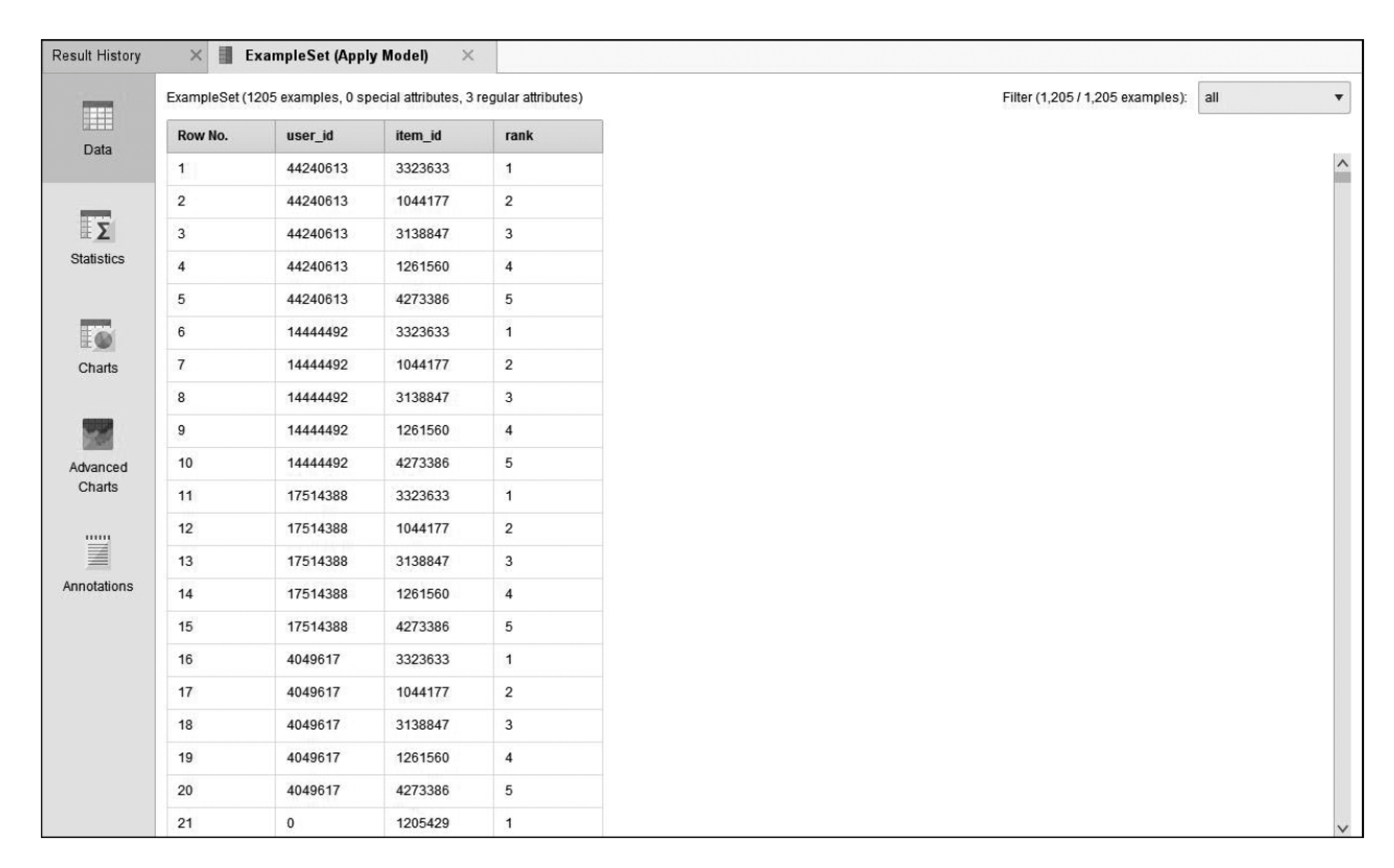


图 6-30　设置推荐商品的数量



Row No.	user_id	item_id	rank
1	44240613	3323633	1
2	44240613	1044177	2
3	44240613	3138847	3
4	44240613	1261560	4
5	44240613	4273386	5
6	14444492	3323633	1
7	14444492	1044177	2
8	14444492	3138847	3
9	14444492	1261560	4
10	14444492	4273386	5
11	17514388	3323633	1
12	17514388	1044177	2
13	17514388	3138847	3
14	17514388	1261560	4
15	17514388	4273386	5
16	4049617	3323633	1
17	4049617	1044177	2
18	4049617	3138847	3
19	4049617	1261560	4
20	4049617	4273386	5
21	0	1205429	1

图 6-31　对用户推荐的商品结果

第 7 章　大数据资源的元数据管理

在大数据应用分析中,分析人员要运用的数据类型繁多、体量庞大、来源杂乱,在分析过程中产生的数字信息也越来越丰富。如何有效地组织、管理和维护海量的数据和信息,以便分析人员访问并综合利用,是一个重要问题。近年来,元数据管理技术日益成熟,元数据作为描述数据的数据的作用已变得越来越重要,成为数据信息资源有效管理和应用的重要手段。充分利用元数据管理技术,结合分析需要制定一套科学的、适用的元数据规范,是十分必要的。本章我们以大数据的审计应用分析为例,紧密结合大数据应用分析的实际需求,探索建立相关内容的元数据规范。

7.1　元数据简介

7.1.1　元数据和对象数据

元数据是英文单词“metadata”的中文意译,若从英文直译则为“关于数据的数据或描述其他信息的数据”。通俗地讲,元数据就是描述数字信息资源特征的数据,它的用途是描述、识别和检索数字信息资源。早在 20 世纪末,元数据的概念和相关工具就已经出现,但限于当时的数据量还不够大,而元数据本身又包含太多的内容,以至于它并未得到充分利用。而在今天看来,元数据正在成为解决诸多数据问题时必须抓住的一个“精髓”要素。

与元数据相对应的一个概念是对象数据。对象数据就是指数字信息资源本身,它可以是以各种形式存在的数字信息,如 Word 文件、Excel 文件、图像、声音和视频等。以图书馆为例,可以将图书馆中的每一本书的正文内容看作对象数据,将书的书名、作者、版本、出版社、出版时间、内容简介和馆藏位置等信息编制成一条卡片目录。这条关于图书的卡片目录的内容就可以称作元数据。显然,有了卡片目录,读者查询图书信息就方便快捷了很多,读者可以在图书馆的卡片目录中查找所需图书的元数据(该图书的书名、作者、版本、出版社、出版时间、内容简介和馆藏位置等信息),然后图书管理员就可以根据读者提供的图书元数据找到读者所需要的对象数据(书籍)。

7.1.2　应用元数据管理技术的意义

首先,大数据资源具有多种多样的格式和控制方式,不容易被人们直接检索。例如,数字资源可能以多种形式存在,既可以是 Word 文档,也可以是社交网页、图像、声音和视频,还可以是卫星、传感器数据。用户可能不太了解和熟悉其中的某种格式的数字资源,因而直接在对象数据中查找所需要的信息会比较困难;另外,数字资源的存取是受控制的,也许它的内容被加密了或者它的内容层层解码、降维以后才能访问,那么在这种情况

下直接检索对象数据也是很困难的。如果设置了元数据来描述对象数据的特征和存放的位置，人们只需以统一的方式在元数据中检索就可以方便迅速地查找到自己需要的对象数据，而不会被对象数据格式的多样性和控制方式所影响。

其次，设置元数据可以提高检索的效率。例如，假设现有 750 万份文献，每份文献有 200 页，每页有 400 个汉字。按一个汉字使用 2 字节计算，如果将这些文献数字化，需要的存储空间为：7 500 000×200×400×2＝1 200GB(GB 代表亿字节)。如果直接在这 1 200GB 的全文中检索我们需要的信息，所使用的时间和检索出的无用的信息都是难以想象的。如果采用元数据方式，假设描述每份文献平均需要 1 500 字节，那么这些文献的元数据的存储空间为：7 500 000×1 500＝11.25GB。毫无疑问，11.25GB 的数据量远远小于 1 200GB，在这 11.25GB 的元数据中检索所需要的文献，与简单的全文检索相比，使用的时间将大为缩减，检索出信息的准确率将得到极大的提高。

最后，通过元数据和对象数据来管理数字资源具有良好的可扩展性。如果把所有的数字资源全部放在一台计算机中供用户使用，那么随着数字资源增加到一定的数量，计算机将由于容量有限而不能保存新增加的数字资源，并且当存在大量的用户让一台计算机在如此众多的数字资源中查找所需要的信息时，计算机检索所使用的时间和检索出的无用的信息都是用户无法接受的。如果通过元数据和对象数据两种方式进行管理，情况将会发生很大的变化。由于元数据的数据量比对象数据的数据量小很多，可以把元数据集中存放在一台计算机中供所有的用户查询使用，而对象数据不必集中保存在一台计算机中，可以充分利用云计算，保存在地理位置分散的多个计算机系统中。用户、保存元数据的计算机和保存对象数据的计算机通过网络连接起来。用户通过网络查询元数据找到对象数据的保存位置，然后再通过网络从保存的位置获得所需要的数字资源。这个过程如图 7-1 所示。

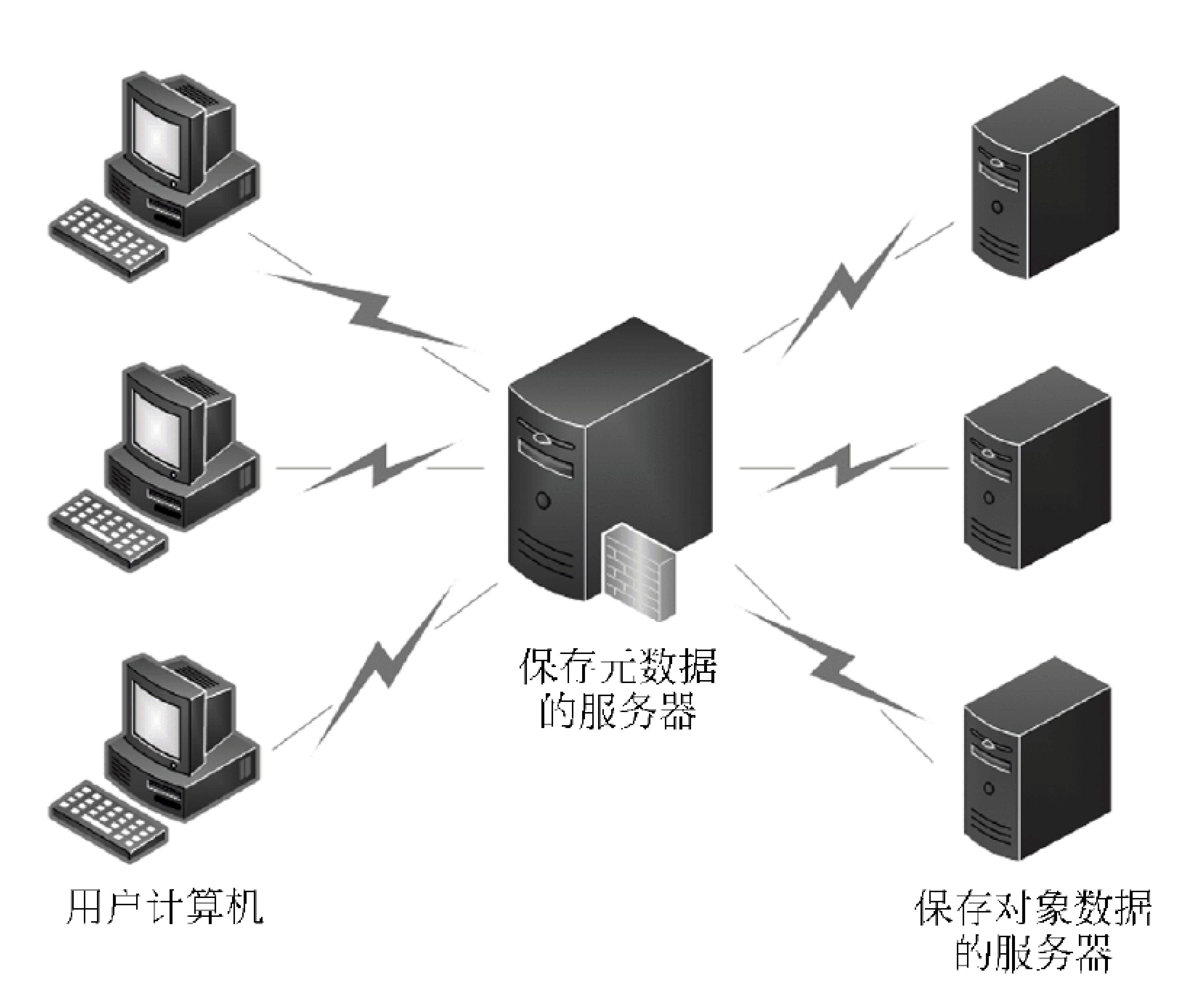


图 7-1　基于元数据和对象数据的组织管理方式

在网络环境中，往往只需几秒钟的时间就能完成上述数字资源的检索和使用。如果某个单位想要增加数字资源，只需把增加的数字资源放入保存对象数据的计算机中，然后

再把相应元数据上传到保存元数据的计算机中，就可以让所有用户查找到新增加的数字资源。只要把地理位置分散的各个单位的元数据放到网络上供用户共享和检索使用，就可以有效地解决各个单位资源利用的关键问题，用户通过检索元数据就能知道谁有什么对象数据，从而有效地提高这些对象数据的利用率。

7.2 著录对象分析

大数据审计分析数字信息是审计数据分析人员在分析过程中产生或利用的数字资源，包含多个内容，本节结合分析业务，仅择要分析其中的少量几项。

7.2.1 审计中间表

1. 定义及特点分析

简单来说，审计中间表是审计人员进行数据分析的对象、资源和平台。它是将转换、清理、验证后的源数据按照提高审计分析效率、实现审计目的的要求进一步选择、整合而形成的数据集合。其特点表现如下。

- 在表现形式上，审计中间表是有着严格创建规范的审计数据，有着较强的业务性，审计人员会根据具体的业务设计较为固定的结构。
- 在内容描述上，审计中间表与被审计单位的生产经营活动密切相关，它随着被审计单位业务量的变化而变化，其内容具有动态性。
- 在文件格式上，采取的是 SQL Server 的数据库格式。SQL Server 数据库有两类文件，分别为数据文件和日志文件。

2. 著录对象范围的界定

从大类上进行区分，审计中间表可划分为基础性中间表和分析性中间表。基础性中间表是审计人员结合被审计单位的业务性质和数据结构，根据不同的分析主题生成的，是面向审计项目组全体审计人员的。例如在海关审计中，基础性中间表不仅包括海关本身的征税、加工贸易、减免税等数据，还包括码头、船舶公司、外汇管理、税收、电子口岸等方面的数据。分析性中间表是审计人员在数据分析过程中，在基础性中间表的基础上根据具体的审计目标和分析需求生成的，它是面向审计组中特定审计人员的。在这里，我们的著录对象既包括基础性中间表，也包括分析性中间表。

3. 著录单位的界定

审计中间表的著录单位是单张表，它是计算机审计过程中产生的最基本的审计数据集合。同一个审计项目会产生多张审计中间表，著录时应把相互之间有包容关系或参照关系的审计中间表当作两个著录单位。此外，应将不同格式的审计中间表当作两个著录单位，即同一内容、不同格式的审计中间表，作为不同对象著录。

7.2.2 审计分析模型

1. 定义及特点分析

审计分析模型是审计人员用于数据分析的数学公式或逻辑表达式，它是按照审计事

项应该具有的性质或数量关系，由审计人员通过设定计算、判断或限制条件建立起来的，用于验证审计事项实际的性质或数量关系，从而对被审计单位经济活动的真实、合法、效益情况做出科学的判断。审计分析模型的特点表现如下。

- 在表现形式上，审计分析模型有多种表现形态：用在查询分析中，表现为一个或一组查询条件；用在多维分析中，表现为切片、切块、旋转、钻取、创建计算成员、创建计算单元等；用在挖掘分析中，表现为设定挖掘条件。
- 在内容描述上，审计分析模型通过审计分析模型算法体现，内容包括构建审计分析模型的思路、方法和步骤，从分析实质上，审计分析模型是一个数学公式或者逻辑表达式。
- 在文件格式上，多维数据集，以各数据库特有的数据集文件格式（如 CAB 或 ABF）存放；数据透视表或图，以 Excel 文件格式（如 XLS 或 XLSX）存放；查询语句，以 SQL 文件格式存放。

2. 著录对象范围的界定

按照在审计中的不同功能，可以将审计分析模型具体划分为系统分析模型、类别分析模型和个体分析模型三大类。系统分析模型主要用于对被审计单位的数据进行整体层次上的全面、系统分析，发现趋势和异常，帮助审计人员把握被审计单位的总体情况。类别分析模型主要按业务类别对审计数据进行分析，指引审计人员发现和锁定重点审计的内容、范围。个体分析模型主要用于核查问题、筛选线索，为延伸取证提供明确具体的目标。

审计人员在某个审计项目中实际构建，经检验能够帮助审计人员验证审计事项实际性质或数量关系，并对被审计单位经济活动的真实、合法及效益情况作出科学判断的审计分析模型，都应该被著录到这个审计项目的审计分析模型数据库中。

3. 著录单位的界定

审计分析模型的著录单位为单个审计分析模型。审计分析模型是通过审计分析模型算法体现的，每个审计分析模型都构建在一个具体的审计项目里。因为被审计单位审计数据结构和数据内容的千差万别和不断变化，所以审计分析模型必须针对具体的审计中间表来构建。在著录审计分析模型时，应注意在审计分析模型的元数据中描述该模型所应用的审计中间表的名称。

7.2.3　审计专家经验

1. 定义及特点分析

审计专家经验是指审计人员在审计分析实践中形成和积累，并已被证明有效的审计知识、技能、方法等。审计专家经验表现如下。

- 在表现形式上，审计专家经验规定了必备的要素，包括标题、经验类别、经验种类、经验描述、审计步骤、经验模型、类 SQL 描述、适用法规、典型案例、资料参数、作者、日期。
- 在内容描述上，审计专家经验是对审计人员在审计实践中形成的审计知识、技能、方法的归纳提炼。
- 在文件格式上，采用的是 PDF 文件格式。

2. 著录对象范围的界定

审计专家经验的著录对象是经各级审计组织征集、评选后确定的优秀审计专家经验。未经甄选的审计专家经验一般不进行著录。这样做的目的主要是保证审计专家经验的质量。

3. 著录单位的界定

审计专家经验的著录单位为单个审计专家经验。

7.2.4 审计情景案例

1. 定义及特点分析

审计情景案例是审计分析过程的情景再现,是以图文形式描述的特定审计情景,一般应包括一个或多个疑难问题,同时也包含解决这些问题的方法。情景案例的特点表现如下。

- 在表现形式上,审计情景案例都规定了必备的要素,包括标题、案例背景、案例实体、案例分析。
- 在内容描述上,审计情景案例是以某个情景为背景,对审计人员在审计实践中形成的审计知识、技能、方法的描述。
- 在文件格式上,采用的是 PDF 文件格式,也可以有图片、音像制品。

2. 著录对象范围的界定

审计情景案例的著录对象是经各级审计组织征集、评选后确定的优秀审计情景案例。未经甄选的审计情景案例一般不进行著录。这样做的目的主要是保证审计情景案例的质量。

3. 著录单位的界定

审计情景案例的著录单位为单个情景案例。

7.2.5 被审计单位资料

1. 定义及特点分析

被审计单位资料是审计分析中收集的对今后审计有指导借鉴意义的被审计单位的相关材料,特点表现如下。

- 在表现形式上,被审计单位资料形式多样,既可以是报告材料,也可以是账表信息。
- 在内容描述上,被审计单位资料内容丰富,包括被审计单位组织沿革、会计资料、审计查出问题描述等。
- 在文件格式上,采用的是 PDF 文件格式和 RAR 文件压缩格式,也可以有图片、音像制品。

2. 著录对象范围的界定

被审计单位资料的著录对象是经审计人员甄别的有审计价值的资料。

3. 著录单位的界定

被审计单位资料的著录单位为单个被审计单位资料或被审计单位资料压缩包。

7.3　元数据结构设计

借鉴《我国数字图书馆标准规范建设：专门数字对象描述元数据规范设计指南》对元数据结构的设计指导意见，根据大数据审计分析的特点和属性，其元数据基本结构由核心元素、审计数字信息核心元素和个别元素三部分组成。

核心元素为各类审计数字信息对象与 DC(Dublin Core，都柏林核心元数据集)保持一致、易于交换的元素。审计数字信息核心元素是除核心元素以外，为某一类审计数字信息的资源对象所通用的元素，如通用公文核心元素包括发文字号、行文依据、行文对象、紧急程度、附件、过程文件。个别元素依据特定使用的资源属性来确定，如审计业务文书中的外资运用审计报告的个别元素就是项目名称、项目执行单位、会计年度。

审计数字信息的元数据著录包括下述各类审计数字信息分别包括的元素及其相应的修饰词。除了必备和有则必备的数据项外，并不一定具备所有的元素和修饰词。本设计不对元数据记录中的各项元素的排列顺序作强制性规定，应用者可以根据用户使用的习惯及其他需求，自行决定元素的排列顺序。

7.3.1　审计中间表的元数据结构

审计中间表著录的元素及元素修饰词包括：标题、标识符、责任者、主题、描述、审计组织、创建日期、资源类型、文件格式、技术环境、语种、控制标识、密级、保密期限、包含于、时间范围、审计项目、表类别、被审计单位、数据库。表 7-1 列出了审计中间表的元数据，表 7-2 列出了审计中间表的元素修饰词及编码体系修饰词。

表 7-1　审计中间表的元数据

核心元素(13 个)			审计中间表核心元素(4 个)	审计中间表个别元素
元素名称	与 DC 的映射(中文)	与 DC 的映射(英文)		
标题	题名	Title	审计项目	根据特定的审计中间表属性来确定
标识符	识别符	Resource Identifier	表类别	
责任者	创建者	Creator	被审计单位	
主题	主题	Subject and Keywords	数据库	
描述	描述	Description		
其他责任者	其他贡献者	Contributor		
日期	日期	Date		
资源类型	类型	Type		
格式	格式	Format		
语种	语言	Language		
权限管理	权限	Rights		
相关信息	关系	Relation		
覆盖范围	覆盖范围	Coverage		

表 7-2　审计中间表的元素修饰词及编码体系修饰词

元素名称	元素修饰词名称	编码体系修饰词
标题		
标识符		URI
责任者		
主题		公文主题词表
描述		
其他责任者	审计组织	
日期	创建日期	Period、W3CDTF
资源类型		指定值“审计中间表”
格式	文件格式、技术环境	
语种		ISO639-2、RFC1766
权限管理	控制标识、密级、保密期限	
相关信息	包含于	
覆盖范围	时间范围	
审计项目		
表类别		
被审计单位		

7.3.2　审计分析模型的元数据结构

审计分析模型著录的元素及元素修饰词包括：标题、标识符、责任者、主题、描述、审计机关、创建日期、资源类型、文件格式、技术环境、语种、控制标识、相关信息、审计项目、模型类别。表 7-3 列出了审计分析模型的元数据，表 7-4 列出了审计分析模型的元素修饰词及编码体系修饰词。

表 7-3　审计分析模型的元数据

<table>
<tr><th colspan="3">核心元素(12 个)</th><th rowspan="2">审计分析模型核心元素(2 个)</th><th rowspan="2">审计分析模型个别元素</th></tr>
<tr><th>元素名称</th><th>与 DC 的映射(中文)</th><th>与 DC 的映射(英文)</th></tr>
<tr><td>标题</td><td>题名</td><td>Title</td><td>审计项目</td><td rowspan="6">根据特定的审计分析模型属性来确定</td></tr>
<tr><td>标识符</td><td>识别符</td><td>Resource Identifier</td><td>模型类别</td></tr>
<tr><td>责任者</td><td>创建者</td><td>Creator</td><td></td></tr>
<tr><td>主题</td><td>主题</td><td>Subject and Keywords</td><td></td></tr>
<tr><td>描述</td><td>描述</td><td>Description</td><td></td></tr>
<tr><td>其他责任者</td><td>其他贡献者</td><td>Contributor</td><td></td></tr>
</table>

续表

核心元素(12 个)			审计分析模型核心元素(2 个)	审计分析模型个别元素
元素名称	与 DC 的映射(中文)	与 DC 的映射(英文)		
日期	日期	Date		
资源类型	类型	Type		
格式	格式	Format		
语种	语言	Language		
权限管理	权限	Rights		
相关信息	关系	Relation		

表 7-4　审计分析模型的元素修饰词及编码体系修饰词

元素名称	元素修饰词名称	编码体系修饰词
标题		
标识符		URI
责任者		
主题		公文主题词表
描述		
其他责任者	审计组织	
日期	创建日期	Period、W3CDTF
资源类型		指定值“审计分析模型”
格式	文件格式、技术环境	
语种		ISO639-2、RFC1766
权限管理	控制标识	
相关信息		
审计项目		
模型类别		

7.3.3　审计专家经验的元数据结构

审计专家经验著录的元素及元素修饰词包括：标题、标识符、责任者、主题、描述、审计组织、创建日期、资源类型、页数、语种、控制标识、相关信息、经验类型、典型案例。

表 7-5 列出了审计专家经验的元数据，表 7-6 列出了审计专家经验的元素修饰词及编码体系修饰词。

表 7-5　审计专家经验的元数据

<table>
<tr><th colspan="3">核心元素(12 个)</th><th rowspan="2">审计专家经验核心元素(2 个)</th><th rowspan="2">审计专家经验个别元素</th></tr>
<tr><th>元素名称</th><th>与 DC 的映射(中文)</th><th>与 DC 的映射(英文)</th></tr>
<tr><td>标题</td><td>题名</td><td>Title</td><td>经验类型</td><td rowspan="12">根据特定的审计专家经验属性来确定</td></tr>
<tr><td>标识符</td><td>识别符</td><td>Resource Identifier</td><td>典型案例</td></tr>
<tr><td>责任者</td><td>创建者</td><td>Creator</td><td></td></tr>
<tr><td>主题</td><td>主题</td><td>Subject and Keywords</td><td></td></tr>
<tr><td>描述</td><td>描述</td><td>Description</td><td></td></tr>
<tr><td>其他责任者</td><td>其他贡献者</td><td>Contributor</td><td></td></tr>
<tr><td>日期</td><td>日期</td><td>Date</td><td></td></tr>
<tr><td>资源类型</td><td>类型</td><td>Type</td><td></td></tr>
<tr><td>格式</td><td>格式</td><td>Format</td><td></td></tr>
<tr><td>语种</td><td>语言</td><td>Language</td><td></td></tr>
<tr><td>权限管理</td><td>权限</td><td>Rights</td><td></td></tr>
<tr><td>相关信息</td><td>关系</td><td>Relation</td><td></td></tr>
</table>

表 7-6　审计专家经验的元素修饰词及编码体系修饰词

元 素 名 称	元素修饰词名称	编码体系修饰词
标题		
标识符		URI
责任者		
主题		公文主题词表
描述		
其他责任者	审计组织	
日期	创建日期	Period、W3CDTF
资源类型		指定值“审计专家经验”
格式	页数	
语种		ISO639-2、RFC1766
权限管理	控制标识	
相关信息		
经验类型		
典型案例		

7.3.4　审计情景案例的元数据结构

审计情景案例著录的元素及元素修饰词包括：标题、标识符、责任者、主题、描述、审计组织、创建日期、资源类型、页数、语种、控制标识、相关信息、案例类型、典型案例。

表 7-7 列出了审计情景案例的元数据，表 7-8 列出了审计情景案例的元素修饰词及编码体系修饰词。

表 7-7　审计情景案例的元数据

核心元素(12 个)			审计情景案例核心元素(2 个)	审计情景案例个别元素
元素名称	与 DC 的映射(中文)	与 DC 的映射(英文)		
标题	题名	Title	案例类型	
标识符	识别符	Resource Identifier	典型案例	
责任者	创建者	Creator		
主题	主题	Subject and Keywords		
描述	描述	Description		
其他责任者	其他贡献者	Contributor		根据特定的审计情景案例属性来确定
日期	日期	Date		
资源类型	类型	Type		
格式	格式	Format		
语种	语言	Language		
权限管理	权限	Rights		
相关信息	关系	Relation		

表 7-8　审计情景案例元素修饰词及编码修饰词

元 素 名 称	元素修饰词名称	编码体系修饰词
标题		
标识符		URI
责任者		
主题		公文主题词表
描述		
其他责任者	审计组织	
日期	创建日期	Period、W3CDTF
资源类型		指定值“审计专家经验”
格式	页数	
语种		ISO639-2、RFC1766

续表

元 素 名 称	元素修饰词名称	编码体系修饰词
权限管理	控制标识	
相关信息		
案例类型		
典型案例		

7.3.5 被审计单位资料的元数据结构

被审计单位资料著录的元素及元素修饰词包括：标题、标识符、责任者、主题、描述、其他责任者、日期、资源类型、资源载体、文件格式、语种、控制标识、密级、保密期限、相关信息、覆盖范围、被审计单位。

表 7-9 列出了被审计单位资料的元数据，表 7-10 列出了被审计单位资料的元素修饰词及编码体系修饰词。

表 7-9 被审计单位资料的元数据

核心元素(13 个)			被审计单位资料核心元素(1 个)	被审计单位资料个别元素
元素名称	与 DC 的映射(中文)	与 DC 的映射(英文)		
标题	题名	Title	被审计单位	根据特定的被审计单位资料属性来确定
标识符	识别符	Resource Identifier		
责任者	创建者	Creator		
主题	主题	Subject and Keywords		
描述	描述	Description		
其他责任者	其他贡献者	Contributor		
日期	日期	Date		
资源类型	类型	Type		
格式	格式	Format		
语种	语言	Language		
权限管理	权限	Rights		
相关信息	关系	Relation		
覆盖范围	覆盖范围	Coverage		

表 7-10 被审计单位资料的元素修饰词及编码体系修饰词

元 素 名 称	元素修饰词名称	编码体系修饰词
标题		
标识符		URI

续表

元素名称	元素修饰词名称	编码体系修饰词
责任者		
主题		公文主题词表
描述		
其他责任者		
日期		Period、W3CDTF
资源类型		指定值“被审计单位资料”
格式	资源载体、文件格式	
语种		ISO639-2、RFC1766
权限管理	控制标识、密级、保密期限	
相关信息		
覆盖范围		
被审计单位		

7.4　应用大数据审计分析数字信息元数据规范的扩展规则

分析人员在应用审计数字信息元数据规范时，可以采取增加元素的方式对大数据审计分析元数据进行扩展。这种扩展方式是横向扩展，应遵守以下规则：

- 按照核心元素、审计数字信息核心元素、个别元素的结构组成。
- 最大可能采用本元数据规范推荐的元数据项，并在语义上严格保持一致。
- 对推荐的元素不能描述的特性可以增加元素，但新增加元素不能与已有元素有任何语义上的重复。

为了更为准确地描述对象，可以使用修饰词的方式向下扩展一层，这是对大数据审计分析数字信息元数据进行纵向扩展。修饰词分为元素修饰词和编码体系修饰词两种，其中元素修饰词(element refinement)是对元素的语义进行修饰，提高元素的专指性和精确性；编码体系修饰词(encoding scheme)则包括控制词表和正规的符号或者解读方式。审计数字信息元数据的纵向扩展应遵守以下规则：

- 如果元素已复用 DC，则修饰词尽量采用 DC 的修饰词。
- 尽可能遵守向上兼容(dump-down)原则，增加的修饰词的语义不能超过被修饰词(元素)的语义，修饰词只是对未修饰词的含义范围做了进一步的限定。
- 修饰词的设定可以复用来自其他元数据标准的修饰词，但要求必须说明来源，使用时严格遵守其语义。
- 自行制订的修饰词必须遵循向上兼容原则，即修饰词的语义包含于相应的未限定元素中，在范围上对未限定元素的语义进行限定，在深度上对未限定元素的语义

进行延伸。对于未具备修饰词知识的用户而言,修饰词可以像未限定元素一样使用。

参考文献

本章在撰写过程中参考了审计署京津冀特派员办事处2007年《审计数字化建设》课题研究成果,课题组主要研究成员有刘汝焯、陈峰、孙俭、严晓健、姚瑜。

第 8 章　大数据分析的数据清洗

8.1　大数据清洗的基本概念

随着信息技术的快速发展，各个领域都以惊人的速度不停地产生各式各样的规模巨大的数据信息，人类也在工作生活的方方面面接触到越来越多的数据信息。然而，人类对数据信息理解的匮乏与数据爆炸的趋势显得并不对称，人类在努力将数据信息转化为有利信息知识的同时，也面临大数据之中夹杂的“脏数据”的挑战。对原始数据源的清洗，将其转化为可被理解利用的目标数据源，成为人类理解数据的过程中尤为重要的一步。

8.1.1　大数据清洗的基本架构

大数据时代，随着“数据驱动运营”的意识在各行业中的逐渐普及，“数据驱动下的精细化运营”也成了生产运营过程中的必然趋势，但在其中同样面临“脏数据”的严峻挑战。

在席卷全球的信息化浪潮中，互联网、云计算、物联网等技术迅猛发展、加速创新，其中积淀的数据爆炸式增长，成为重要的生产要素和社会财富，堪称信息时代的矿产和石油。针对这种史无前例的数据洪流，如何挖掘信息时代的“数字石油”，将大数据转换为大智慧、大市场和大生态，是我们这个时代的历史机遇。

大数据已经渗透到各个行业和业务职能领域，成为重要的生产因素。数据的来源主要有政府数据、行业数据、企业数据和从数据交易所交换的数据。

海量数据的不断剧增形成大数据时代的显著特征，而大数据的生产和交易的重要前提之一是数据的清洗。大数据必须经过清洗、分析、建模、可视化才能体现其潜在的价值。但由于业务应用的多样化和社交网络的繁荣，单个文件(如日志文件、音视频文件等)变得越来越大，硬盘的读取速度越来越无法满足人们的需要，文件的存储成本越来越高。与此同时，政府、银行和保险公司等内部存在海量的非结构化、不规则的数据，只有将这些数据采集并清洗为结构化、规则的数据，才能提高部门的决策支撑能力和政府决策服务水平，使之发挥应有的作用。

因此，目前的数据清洗主要是将数据划分为结构化数据和非结构化数据，分别采用传统的数据提取、转换、加载(ETL)工具和分布式并行处理来实现，其基本架构如图 8-1 所示。

结构化数据一般存储在传统的关系型数据库中。关系型数据库在处理事务、及时响应、保证数据的一致性方面有天然的优势。

非结构化数据可以存储在新型的分布式存储中，如 Hadoop 的 HDFS。分布式存储在系统的横向扩展性、降低存储成本、提高文件读取速度方面有着独特的优势。

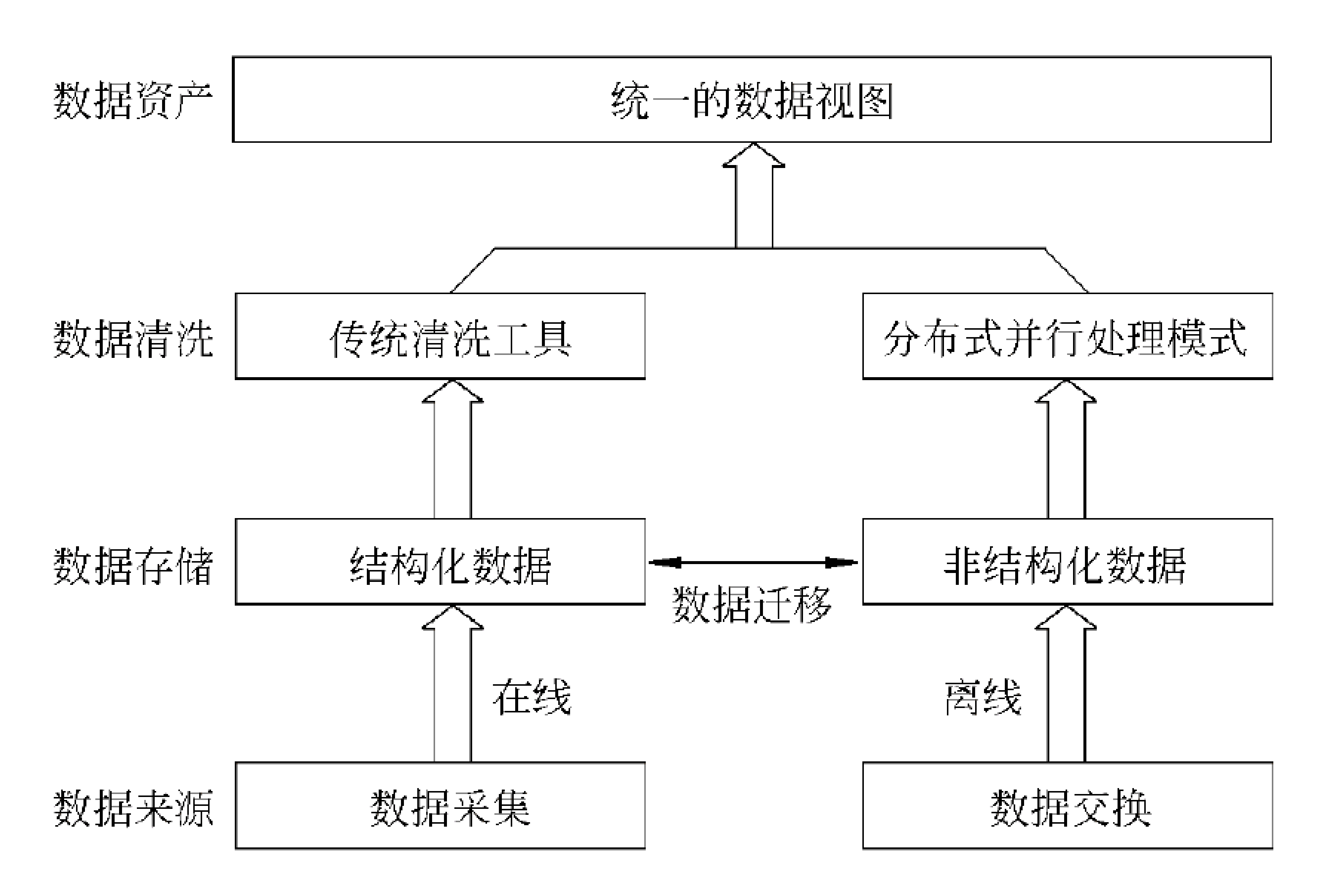


图 8-1　大数据清洗的基本架构

数据清洗在汇聚多个维度、多个来源、多种结构的数据之后，就可以对数据进行抽取、转换和集成加载。在这个过程中，除了更正、修复系统中的一些错误数据之外，更多的是对数据进行归并整理，并储存到新的存储介质中。其中，分清和掌握数据的质量至关重要。

常见的数据质量问题可以根据数据源的多少和所属层次(定义 Scheme 层和实例 sample 层)分为以下四类。

第一类，单数据源定义层：违背字段约束条件(如日期出现 1 月 0 日)、字段属性依赖冲突(如两条记录描述同一个人的某一个属性，但数值不一致)、违反唯一性(同一个主键 ID 出现了多次)。

第二类，单数据源实例层：单个属性值含有过多信息、拼写错误、空白值、噪声数据、数据重复、过时数据等。

第三类，多数据源的定义层：同一个实体的不同称呼(如冰心和谢婉莹，用笔名还是用真名)、同一种属性的不同定义(如字段长度定义不一致、字段类型不一致等)。

第四类，多数据源的实例层：数据的维度、粒度不一致(如有的按 GB 记录存储量，有的按 TB 记录存储量；有的按年度统计数据，有的按月份统计数据)、数据重复、拼写错误等。

除此之外，还有在数据处理过程中产生的“二次数据”，其中也会有噪声、重复或错误的情况。数据的调整和清洗也会涉及格式、测量单位和数据标准化与归一化的相关事情。通常这类问题可以归结为不确定性。不确定性有两方面内涵，包括各数据点自身存在的不确定性，以及数据点属性值的不确定性。前者可以用概率描述，后者有多重描述方式，如描述属性值的概率密度函数，以方差为代表的统计值等。

8.1.2　数据清洗的基本步骤

数据清洗指的是把“脏”数据“洗掉”，即发现并纠正数据中可识别的错误，包括检查数据一致性、处理无效值和缺失值等。在大数据时代，数据的种类和来源众多，这就避免不

了有错误的数据或异常的数据，这些错误的数据称为“脏数据”。我们要按照一定的规则把“脏数据”洗掉，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果报给相关部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。

数据清洗是整个数据分析过程中不可或缺的一个环节，数据清洗的质量直接关系到模型效果和最终结论。在实际操作中，数据清洗通常会占用分析过程的 50%～80%的时间。

无论用海量数据还是大数据来表征这个时代，数据规模庞大、增长迅速、类型繁多、结构各异已成为无法回避的现实问题。如何把繁杂的大数据变成我们能应付的、有效的“小”数据，即针对特定问题构建一个干净、完备的数据集，这一过程变得尤为重要。在大数据时代，若不加强数据清洗，则 GIGO(垃圾进，垃圾出)现象会更加严重。

对数据清洗之后进行分析挖掘的过程就是“去粗取精、去伪存真、化零为整、见微知著”的过程。只有通过清洗与过滤得到干净完备的数据，才能通过分析与挖掘得到可以让人放心的、可用于支撑决策的信息。

在进行数据分析之前，首先应该进行数据清洗，在开始数据清洗之前，应先对数据进行必要的预处理。

数据预处理阶段主要做两件事情：一是将数据导入处理工具；二是查看数据，包括查看元数据，查看字段解释、数据来源、代码表等一切描述数据的信息，另外还需抽取一部分数据，使用人工查看方式，对数据本身有一个直观的了解，并且初步发现一些问题，为之后的处理做准备。

数据预处理之后，就可以进行数据清洗了。数据清洗通常包含如下几个步骤。

1. 缺失值清洗

缺失值是最常见的数据问题，处理缺失值也有很多方法，一般建议按照以下四个步骤进行。

(1) 缺失值

对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定不同的策略。

- 对于重要性高、缺失率低的情况，可以通过估值计算的方法进行填充，或者通过经验或业务知识进行估计。
- 对于重要性高、缺失率也高的情况，可尝试从其他渠道补全数据，或者使用其他字段通过计算获取缺失值，如果该字段对分析影响很小也可以直接去除该字段；
- 对于重要性低、缺失率低的情况，可以不进行处理或只进行简单填充；
- 对于重要性低、缺失率高的情况，可以去除该字段。

(2) 去除不需要的字段

去除不需要的字段很简单，只要直接删除即可，但在删除之前应该对所做的每一步清洗操作都进行记录，并对数据进行备份。

(3) 填充缺失内容

对可以进行填充的缺失值，可采用如下方法进行填充：

- 以业务知识或经验推测填充缺失值；
- 以同一指标的计算结果(如均值、中位数、众数等)填充缺失值；

- 以不同指标的计算结果填充缺失值，例如，假设年龄段有缺失值，但有身份证号信息，则可利用身份证号来填充年龄的缺失值。

(4) 重新取数

如果某些指标非常重要且缺失率高，则需要向有关业务人员进行了解，是否有其他渠道可以取得相关数据。

2. 格式内容清洗

如果数据是由系统日志而来，那么通常在格式和内容方面会与元数据的描述一致。但如果数据是由人工收集或用户填写而来，则有很大可能性在格式和内容上存在一些问题。简单来说，格式内容问题有以下几类。

(1) 时间、日期、数值、全半角等显示格式不一致

这种问题通常与输入端有关，在整合多来源数据时也有可能遇到，将其处理成一致的某种格式即可。

(2) 内容中有不该存在的字符

某些内容可能只包括一部分字符，如身份证号是数字+字母，中国人姓名是汉字。最典型的就是头、尾、中间的空格，也可能出现姓名中存在数字符号、身份证号中出现汉字等问题。这种情况下，需要以半自动校验半人工方式来找出可能存在的问题，并去除不需要的字符。

(3) 内容与该字段应有内容不符

姓名写了性别、身份证号写了手机号等，均属这种问题。但该问题的特殊性在于：并不能简单地通过删除来处理，因为有可能是人工填写错误，也有可能是前端没有校验，还有可能是导入数据时部分或全部存在列没有对齐的问题，因此要详细识别问题类型。

格式内容问题是比较细节的问题，但很多分析失误都是因为这个问题，如跨表关联失败(多个空格导致工具认为“张长江”和“张 长江”不是一个人)、统计值不全(数字里掺了字母，在进行求和时肯定出问题)、模型输出失败或结果不对(如数据对错列了，把日期和年龄搞混了)。因此，当处理的数据是人工收集而来时，或者当产品前端校验设计的不太好时，必须注意这些数据的清洗问题。

3. 有逻辑错误的数据的清洗

这部分的工作是去掉一些使用简单逻辑推理就可以直接发现问题的数据，防止分析结果走偏。主要包含以下几个步骤。

(1) 去重

例如，由于人工录入等问题，有可能将“ABC 管家公司”误录入为“ABC 官家公司”，如果不仔细看，是看不出两者的区别的，而且就算看出来了，也不能保证没有“ABC 官家公司”的存在。这种情况，只能或者是写模糊匹配算法，或者直接肉眼观察。

(2) 去除不合理值

例如，人的年龄填写为 200 岁，人的性别填写为“男”“女”之外的其他值。这种数据要么删掉，要么按缺失值处理。

(3) 修正矛盾内容

有些字段是可以相互验证的，例如，身份证号是 1101031980××××××××，而年

龄填 18 岁。这种情况，需要根据字段的数据来源判定哪个字段提供的信息更为可靠，去除或重构不可靠的字段。

逻辑错误除了以上列出的情况外，还有其他很多情况，在实际操作中要酌情处理。另外，这一步骤在之后的数据分析建模过程中有可能重复，因为即使问题很简单，也并非所有问题都能够一次找出，我们能做的是使用工具和方法，尽量减少问题出现的可能性，使分析过程更为高效。

4. 非需求数据的清洗

这一步说起来非常简单：把不要的字段删了。但实际操作起来，还是会有很多问题，例如：

- 把看上去不需要但实际上对业务很重要的字段删了；
- 某个字段觉得有用，但又没想好怎么用，不知道是否该删；
- 一时疏忽，删错字段了。

前两种情况的建议是：如果数据量没有大到不删字段就没办法处理的程度，那么能不删的字段尽量不删。对于第三种情况，应该在每次删除字段前备份数据，从而避免因误删字段而丢失数据的情况。

5. 关联性验证

如果数据有多个来源，那么有必要进行关联性验证。例如，假设商品的销售有实体店销售、网上销售等多种渠道。要了解某商品的销售情况，就需要将多种销售渠道销售的商品通过商品号和商品名进行关联，查看不同渠道销售的相同商品是否商品号一致，如果不一致会直接影响数据分析结果。

8.2 数据清洗

本节以清理结构化数据为例，介绍一些常用的数据清洗方法。

8.2.1 数据清洗的一些注意事项

1. 数据备份

由于在清洗过程中，会对数据进行大量修改，为了保证在需要时随时能够得到、对照或恢复清洗前的数据，在清洗前，应对数据进行备份。

2. 谨慎进行清洗

数据清洗应谨慎进行，以免丢失分析线索。应该在清洗前对空值、不合常理的数值等情况进行具体分析，未准确把握其经济含义前不能一概替换为 0 或进行其他更正；对冗余数据的处理也应小心对待，因为如果被分析单位的数据库较规范，一般就不会出现冗余数据，数据库中的数据一般都是有意义的。在清理的过程中，对能准确判断为无经济含义、与数据分析无关的数据，才可以进行删除。对于重复记录，如果存在少量的完全重复的记录，应检查是否存在人为虚增业务量等情况；如果存在大量重复记录，则应考虑是否存在病毒、文件传输或复制中的问题。

示例：可整列删除的情况。

某表的结构在数据字典中的描述如表 8-1 所示。

表 8-1 数据表结构示例

字段名	中文名	类　型	说　明
zkzh	主卡账号	char(16)	not null
khrq	开户日期	date	not null
yslxye	应收利息余额	dec (16,2)	当前透支周期内首次发生应收利息的日期
zdtzlx	最大透支利息	dec(16,2)	历史最大值
ljtzlx	累计透支利息	dec(16,2)	历史累计值
scfsr	上次发生日	date	该账户最后一次发生应收利息的日期
je1	金额 1	dec(16,2)	保留
je2	金额 2	dec(16,2)	保留
dac	数据校验码	char(16)	存放 DAC 值

通过阅读表 8-1,我们很容易判断出该表的最后一列是没有经济含义的,对于数据分析而言是没有实际意义的,对于此种情况的列,可以完全视为冗余列,在清洗时将该列全部删除。

8.2.2 常见的数据清洗

1. 处理冗余数据

(1) 处理重复行数据

如图 8-2 所示为有重复行数据的一个贷款利率表数据示例。如果要保留原始表中的数据,并将删除了重复行的数据保存到另一个新表中(如新贷款利率表),则可使用如下 SQL 语句:

```
SELECT DISTINCT *  INTO 新贷款利率表 FROM 贷款利率表
```

消除了重复行数据后的“新贷款利率表”数据示例如图 8-3 所示。

(2) 处理列中冗余数据

那些对分析来说是多余的数据,可以使用如下语句处理。

示例:消除 X 表中的列 COL1 中所有以“?”开始的数据。

```
Delete from X where COL1 like 's?%'  escape 's'
```

(3) 处理冗余字段

冗余字段是数据分析人员在分析数据中不需要的字段,如图 8-4 所示。这些字段的存在可能会对审计分析人员分析数据造成不必要的麻烦,因此,在对数据进行分析前,可以先将这些冗余字段清除。

这种情况可以使用 ALTER TABLE 语句将冗余字段删除。

示例:删除“分录明细表”中的“职员”“自定义项目”和“汇率”三个列。

	行号	代码	名称	利率	开始日期	总额
1	500	100	90年五年期财政债券利率	8.3100004	1997-01-10 00:00:00	NULL
2	500	1100	单位活期存款	1.2	1998-07-01 00:00:00	NULL
3	500	1100	单位活期存款	1.2	1998-07-01 00:00:00	NULL
4	500	1100	单位活期存款	1.2	1998-07-01 00:00:00	NULL
5	500	1100	单位活期存款	1.2	1998-07-01 00:00:00	NULL
6	500	1101	单位活期存款(协定)	1.38	1998-12-07 00:00:00	NULL
7	500	1200	单位活期存款	6.3000002	1997-10-23 00:00:00	NULL
8	500	1200	单位活期存款	6.3000002	1997-10-23 00:00:00	NULL
9	500	1201	同业存款	5.8499999	1997-10-23 00:00:00	NULL
10	500	200	开发银行债券利率(12.5%)	10.41	1997-01-10 00:00:00	NULL
11	500	200	开发银行债券利率(12.5%)	10.41	1997-01-10 00:00:00	NULL
12	500	201	开发银行债券利率(14%)	11.64	1997-01-10 00:00:00	NULL
13	500	300	总行金融债券利率(11.5%)	9.5699997	1997-01-10 00:00:00	NULL
14	500	301	总行金融债券利率(11.25%)	9.375	1997-01-10 00:00:00	NULL

图 8-2　包含重复行数据的情况

	行号	代码	名称	利率	开始日期	总额
1	500	100	90年五年期财政债券利率	8.31	1997-01-10 00:00:00	NULL
2	500	1100	单位活期存款	1.2	1998-07-01 00:00:00	NULL
3	500	1101	单位活期存款(协定)	1.38	1998-12-07 00:00:00	NULL
4	500	1200	单位活期存款	6.3	1997-10-23 00:00:00	NULL
5	500	1201	同业存款	5.85	1997-10-23 00:00:00	NULL
6	500	200	开发银行债券利率(12.5%)	10.41	1997-01-10 00:00:00	NULL
7	500	201	开发银行债券利率(14%)	11.64	1997-01-10 00:00:00	NULL
8	500	300	总行金融债券利率(11.5%)	9.57	1997-01-10 00:00:00	NULL
9	500	301	总行金融债券利率(11.25%)	9.375	1997-01-10 00:00:00	NULL

图 8-3　消除掉重复行数据后的情形

职员	自定义项目	币别	汇率	原币金额	借方金额
NULL	NULL	RMB	NULL	3962.5	3962.5
NULL	NULL	RMB	NULL	300000.0	300000.0
NULL	NULL	RMB	NULL	14000.0	14000.0
NULL	NULL	RMB	NULL	1000.0	1.0
NULL	NULL	RMB	NULL	20000.0	20000.0
NULL	NULL	RMB	NULL	400.0	1.0
NULL	NULL	RMB	NULL	38000.0	1.0
NULL	NULL	RMB	NULL	2000.0	1.0
NULL	NULL	RMB	NULL	2000.0	1.0
NULL	NULL	RMB	NULL	5000.0	1.0
NULL	NULL	RMB	NULL	8094.9000000000005	1.0
NULL	NULL	RMB	NULL	7655.7799999999997	1.0
NULL	NULL	RMB	NULL	75.0	75.0
NULL	NULL	RMB	NULL	500.0	500.0
NULL	NULL	RMB	NULL	1020.0	1.0

图 8-4　有冗余字段的情形

```
alter table 分录明细表 drop column 职员,自定义项目,汇率
```

2. 处理空值

在对含有空值的列进行分析统计时,可能会对某些结果产生影响,因此,在分析处理数据之前,可以先对这些空值进行处理。例如,对如图 8-5 所示的数据,希望对借方金额、贷方金额列中的空值进行处理,如将这些空值全部替换为 0,可以通过如下 SQL 语句将"借方金额"列和"贷方金额"列的空值全部替换为 0。

```
update 分录明细表 set 借方金额 =0 where 借方金额 is null
update 分录明细表 set 贷方金额 =0 where 贷方金额 is null
```

处理后的结果如图 8-6 所示。

币别	汇率	原币金额	借方金额	贷方金额
RMB	NULL	3962.5	3962.5	NULL
RMB	NULL	760.0	760.0	NULL
RMB	NULL	3962.5	NULL	3962.5
RMB	NULL	760.0	NULL	760.0
RMB	NULL	300000.0	300000.0	NULL
RMB	NULL	300000.0	NULL	300000.0
RMB	NULL	14000.0	14000.0	NULL
RMB	NULL	2000.0	NULL	2000.0
RMB	NULL	1200.0	NULL	1200.0
RMB	NULL	400.0	NULL	400.0
RMB	NULL	1000.0	NULL	1000.0
RMB	NULL	400.0	NULL	400.0
RMB	NULL	400.0	NULL	400.0
RMB	NULL	400.0	NULL	400.0
RMB	NULL	600.0	NULL	600.0
RMB	NULL	3000.0	NULL	3000.0

图 8-5　含有空值的列

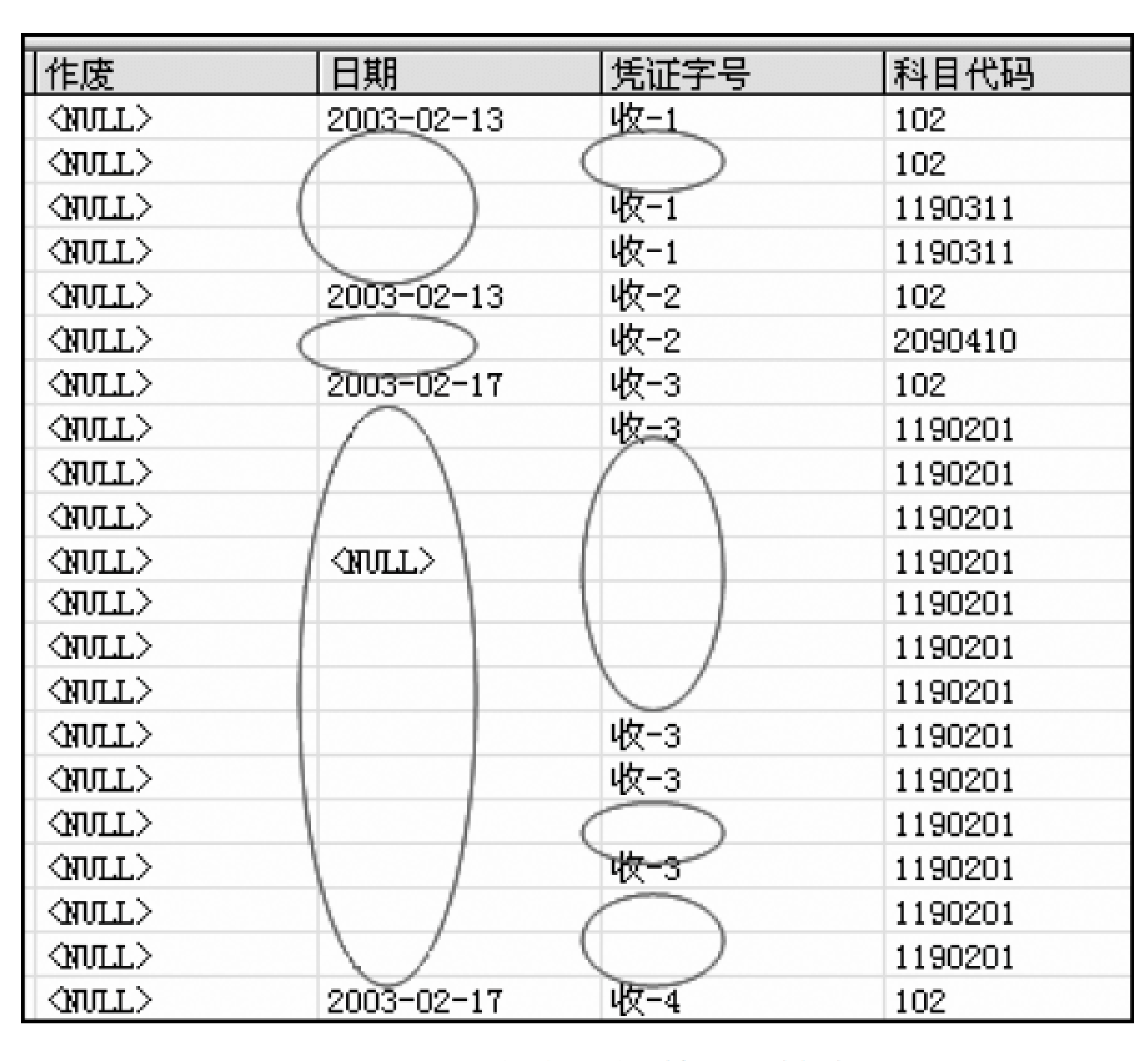
币别	汇率	原币金额	借方金额	贷方金额
RMB	NULL	3962.5	3962.5	0.0
RMB	NULL	760.0	760.0	0.0
RMB	NULL	3962.5	0.0	3962.5
RMB	NULL	760.0	0.0	760.0
RMB	NULL	300000.0	300000.0	0.0
RMB	NULL	300000.0	0.0	300000.0
RMB	NULL	14000.0	14000.0	0.0
RMB	NULL	2000.0	0.0	2000.0
RMB	NULL	1200.0	0.0	1200.0
RMB	NULL	400.0	0.0	400.0
RMB	NULL	1000.0	0.0	1000.0
RMB	NULL	400.0	0.0	400.0
RMB	NULL	400.0	0.0	400.0
RMB	NULL	400.0	0.0	400.0
RMB	NULL	600.0	0.0	600.0
RMB	NULL	3000.0	0.0	3000.0

图 8-6　处理完空值后的情形

3. 处理不规范数据

（1）字段缺失

在录入数据时，操作人员在对多条连续记录中存在的相同数据值进行录入时，可能只录入了第一条记录的数据值，而省略了后续记录的相同数据值的录入，因而导致数据不完整、某些记录有缺失值存在，如图 8-7 所示。在图 8-7 中，“日期”列和“凭证字号”均有缺



作废	日期	凭证字号	科目代码
<NULL>	2003-02-13	收-1	102
<NULL>			102
<NULL>		收-1	1190311
<NULL>		收-1	1190311
<NULL>	2003-02-13	收-2	102
<NULL>		收-2	2090410
<NULL>	2003-02-17	收-3	102
<NULL>		收-3	1190201
<NULL>			1190201
<NULL>			1190201
<NULL>	<NULL>		1190201
<NULL>			1190201
<NULL>			1190201
<NULL>			1190201
<NULL>		收-3	1190201
<NULL>		收-3	1190201
<NULL>			1190201
<NULL>		收-3	1190201
<NULL>			1190201
<NULL>			1190201
<NULL>	2003-02-17	收-4	102

图 8-7　有缺失值的数据

失值。这些不完整的缺失值数据有可能影响数据分析人员对这些数据的分析结果。因此在对数据进行分析之前，需要先对这些缺失值进行处理。

处理缺失值可以使用 SQL Server 的游标机制实现。

示例：处理“车购费数据库”中“分录明细表”中“凭证字号”字段的缺失值。

```
--处理分录明细表的“凭证字号”列的缺失值的代码
--为表增加一个标识列
ALTER TABLE 分录明细表 add tid bigint identity(1,1) not null
  GO
  declare @ id1 bigint,@ id2 bigint
  declare @ value1 nvarchar(25),@ value2 nvarchar(25)    --缺失值列的数据类型
   DECLARE tab_cur cursor for select tid, 凭证字号 from 分录明细表
   OPEN tab_cur
   FETCH NEXT FROM tab_cur into @ id1,@ value1
   FETCH NEXT FROM tab_cur into @ id2,@ value2
        --处理最开始的缺失值
   if @ value1 is null  or @ value1=''
   begin
   --首先找到第一个不为空的值
     while @ @ FETCH_STATUS = 0 and (@ value2 is null  or @ value2='')
     begin
        set @ value1 =@ value2
        set @ id1 =@ id2
        FETCH NEXT FROM tab_cur into @ id2,@ value2
       end
     --然后对开始的这些缺失值进行处理
     while (@ id1 > 0) and (@ @ FETCH_STATUS = 0)
     begin
        update  分录明细表 set 凭证字号 =@ value2 where tid =@ id1
        set @ id1 =@ id1 - 1
    end
       set @ value1 =@ value2
    set @ id1 =@ id2
    FETCH NEXT FROM tab_cur into @ id2,@ value2
   end
   --处理后边的缺失值
   WHILE @ @ FETCH_STATUS = 0
   BEGIN
     --处理连续缺失值情况
     while (@ @ FETCH_STATUS = 0) and (@ value2 is null  or @ value2=''  )
     begin
        print @ value2
        update 分录明细表 set 凭证字号 =@ value1 where tid =@ id2
        FETCH NEXT FROM tab_cur into @ id2,@ value2
```

```
        end
        set @ value1 = @ value2
        set @ id1 = @ id2
        FETCH NEXT FROM tab_cur into @ id2,@ value2
    END
CLOSE tab_cur
DEALLOCATE tab_cur
ALTER TABLE 分录明细表 drop column tid
GO
```

对“分录明细表”处理缺失值后的情况如图 8-8 所示。

作废	日期	凭证字号	科目代码
<NULL>	2003-02-13	收-1	102
<NULL>	2003-02-13	收-1	102
<NULL>	2003-02-13	收-1	1190311
<NULL>	2003-02-13	收-1	1190311
<NULL>	2003-02-13	收-2	102
<NULL>	2003-02-13	收-2	2090410
<NULL>	2003-02-17	收-3	102
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-3	1190201
<NULL>	2003-02-17	收-4	102

图 8-8　处理完缺失值后的情形

（2）无用空格

数据前的无用空格会影响分析人员在按条件进行查询时的查询结果。例如，图 8-9 所示的数据，如果审计分析人员要查看某局报销的卫生费情况，如果在条件语句中写成：WHERE ZY='某局卫生费'，则图 8-9 中的一些数据将不在查询结果中（因为前边有空格），因而影响了审计数据分析的准确性。为了避免这种情况，在对数据进行审计分析之前应先将数据前的这些无用空格去掉。

PZLB	KMBH	ZY	JSH
01	521017	李某医药费	<NULL>
01	102003	李某医药费	5308
01	405004006	某公司的修理费	<NULL>
01	521008020	某公司的修理费	<NULL>
01	102003	某公司的修理费	1365
01	521005003	张某差旅费	<NULL>
01	102003	张某差旅费	5301
01	521006	某局卫生费	<NULL>
01	102003	某局卫生费	1366
01	521006	某局卫生费	<NULL>
01	102003	某局卫生费	1367
01	214001	王某医药费	<NULL>
01	102003	王某医药费	5312

图 8-9　数据前有多余空格的情况

示例：消除“车购费数据库”中“费用记录表”中“ZY”列前的无用空格，可使用如下页所示的 SQL 语句实现：

```
Update 费用记录表 set ZY = LTRIM(ZY)
```

(3) 异常取值数据

异常取值数据是指数据的值存在超出数据字典规定的值域，或与经济含义不相符合等情况的数据。异常取值数据的出现有可能是数据库在取值约束设计上存在问题或缺陷，约束无效、作用弱，导致数据取值有不合常理的情况；另外，异常取值数据的存在也可能表明组织在基础数据的输入中存在失误或舞弊，导致出现与经济活动事实不相符合的数据。

对于异常取值的数据，在清理的过程中一般不应当直接改正或删除，应采取记录或单独保存等谨慎的处理方法，并将数据反映的问题作为待核查的问题。

示例：在对某市公路建设资金的审计过程中，审计人员取得了对某市 3 万多辆国产车的车辆购置附加费的征收情况的数据，在对数据的清理过程中，审计人员发现部分国产车的发票价超过 100 万元，最高的竟高达 740 余万元，如此高的价格明显与现实不符。为了核实是数据录入错误还是将进口车按国产车征收费用，审计人员采取了将发票价超过 100 万元的数据单独保存的方法进行数据清洗。检索出的该部分数据如图 8-10 所示。

档案号	车主名称	发票价	最低征收额	实际征收额	征费日期
	某某市公路管理局	7455000	0	88400	2000-3-6
06009812311008	某某威	5000000	0	80000	1998-12-31
06000001191006	某某市电力局	4000000	0	31790	2000-1-19
06009611221007	某某军	2890000	0	289000	1996-11-29
	某某克斯(某某)有	1170000	110000	117000	1999-6-8
06009906081004	某某克斯(某某)有	1170000	0	117000	1999-7-20
06009905181022	某某克斯(某某)有	1170000	0	117000	1999-7-20
	某某克斯(某某)有	1170000	110000	117000	1999-5-18
	某某海	1045000	0	3000	1998-4-24
06009908101007	某某克斯(某某)有	1000000	100000	100000	1999-8-10
06009907201010	某某克斯(某某)有	1000000	100000	100000	1999-7-20
06009907201012	某某克斯(某某)有	1000000	100000	100000	1999-7-20
06009907201013	某某克斯(某某)有	1000000	100000	100000	1999-7-20

图 8-10　车购费数据清洗中单独保存数据

对存在这些问题的数据要认真进行分析，查明原因后，才能分别情况进行更改处理。

示例：在固定资产表中审计人员发现资产原值字段存在负值的情况，经复核转换的记录，发现是转换人员在转换的过程中将借方金额设定为负、贷方设定为正，导致与审计人员熟悉的以正数表示固定资产原值的规定不符合，针对这种情况，可用如下 SQL 语句处理：

```
Update table set column = abs(column) where column < 0
```

参考文献

[1] [美]Megan Squire. 干净的数据 数据清洗入门与实践[M]. 任政委，译. 北京：人民邮电出版社，2016.

[2] 国研网. 大数据时代亟需强化数据清洗环节的规范和标准[EB/OL]，2015-10-10.

[3] 刘汝焯. 计算机审计质量控制模型：第 2 版[M]. 北京：清华大学出版社，2016.

[4] 中国智能化网. 大数据处理及采集方法，2014-03-24.

[5] 李志刚，等. 大数据——大价值、大机遇、大变革[M]. 北京：电子工业出版社，2012.

第9章　大数据分析的风险与对策

信息化和网络化的高速发展使大数据分析在各行各业都获得了广泛应用。大数据分析在提高经济和社会效益的同时，也面临一定的风险，具体表现为：大数据分析的结果不正确；大数据分析得出的结论不完备；大数据在采集、存储或分析的过程中，数据被非法添加、修改和删除，从而导致数据质量下降，分析结果的可信度降低甚至出现错误；大数据分析的结果泄露个人隐私；大数据被黑客窃取，从而泄露隐私数据等。只有充分认识大数据分析面临的这些风险并采取相应的对策，才能充分发挥大数据分析的作用。本章首先介绍产生大数据分析风险的原因，然后进一步说明大数据在采集、处理与集成、分析方面存在的风险，最后详细叙述大数据分析过程中安全和隐私保护方面面临的风险及其对策。本章的目的是让读者对大数据分析中存在的风险做到“知己知彼”，从而防范和控制风险，保证大数据的正常应用。需要说明的是，从整体来讲，目前大数据风险的对策还不成熟，远远不能满足实际应用的需要。有的对策仍处于研究之中，有的技术和对策虽然初现雏形，但存在适用范围和前提条件，只能应用在某些风险的防范之中。

9.1　大数据分析的风险及产生原因

大数据分析是指对规模巨大的海量数据进行分析，从中寻找模式、相关性和其他有用的信息，帮助用户更好地适应变化，做出更明智的决策。大数据分析的风险并非仅仅局限于数据分析和知识发现的阶段，数据质量、采集方法、数据的处理与集成以及数据解释都会直接或间接影响大数据分析的结果，从而导致风险的发生。

大数据分析的风险主要来自以下三个方面。

1. 大数据固有的复杂性

高质量的数据是大数据分析真实可靠的首要条件，然而大数据的复杂性导致数据质量难以管控，给后续的数据采集与过滤，数据清理与集成以及数据分析带来巨大的挑战，是导致大数据分析风险的最主要原因。大数据的复杂性体现在四个方面。首先，大数据的数据量非常巨大，通常是 TB 或 PB 数量级，而且这些数据关联关系复杂，质量良莠不齐；其次，大数据种类繁多，结构复杂，既包含了传统的结构化数据，又包含了越来越多的文本、图像、声音等半结构化数据和非结构化数据；再次，大数据来源复杂，数据格式千差万别，既有海量的交易数据，又包含海量的网络信息和传感数据；最后，大数据价值密度低，数据量越大，里面真正有价值的东西所占的比例就会越少，大数据分析就像“大海捞针”。

2. 大数据分析的复杂性

目前虽然出现了很多基于大数据分析的成功案例，但是现阶段大数据分析的技术和方法还处于“成长期”，远远不能满足生产和生活中的实际要求，面临许多挑战。例如，大

数据的价值在于从海量数据中挖掘出有用的信息，这依赖于高精度、高效率的机器学习算法来对人类难以理解的底层数据特征进行分析和挖掘。在大数据时代，半结构化和非结构化数据量的迅猛增长，难以采用传统针对结构化数据的分析方法发现其内部关系；同时，随着时间的流逝，大数据中所蕴含的知识价值随之递减，实时处理成为大数据分析的典型需求。在大数据时代，更多应用场景的数据分析从离线分析转向在线实时分析。当很多数据洪流源源不断地涌现时，目前的大数据分析方法还不能有效地对其中数量庞大的半结构化数据和非结构化数据进行实时深度分析。

3. 大数据面临的隐私和安全风险

安全和隐私是信息化社会永恒的主题。在大数据时代，越来越多的数据以数字化的形式存储在电脑中，互联网的发展则使数据更加容易产生和传播，数据安全和隐私问题变得越来越严重，给大数据分析带来严峻的挑战，表现在如下几个方面。

（1）破坏数据质量，影响分析结果的可信度

大数据来源广泛，结构复杂，数据质量难以管控，黑客能够在采集的数据源中通过伪造数据、修改数据等方式，破坏数据的正确性、真实性和完备性，导致数据分析结果可信度下降甚至出现错误的分析结果。

（2）窃取数据，导致隐私泄露

大数据的分析处理流程包括数据采集、存储、传输、集成与处理、分析和解释等环节，敏感的数据（包括分析的结果）在经过这些环节时，可能被黑客窃取，从而导致隐私泄露。例如，在基因大数据分析中，一个人的基因序列属于敏感数据，因为基因序列能够反映其种族、家族、性别、年龄、头发颜色、皮肤颜色、眼睛颜色以及患某种疾病的风险等特征。这些基因数据在传输和存储的过程中，如果不加以保护，就容易被黑客窃取，进而识别出属于这个基因序列的个人及其特征，从而导致个人隐私泄露。国外曾经有过报道，某个基因研究小组收集了 10 万个志愿者的基因序列。尽管该小组把志愿者的姓名、出生年月、邮政编码和性别作了匿名处理，但是把这些基因序列数据和公民信息进行融合后，最终还是甄别出 84％～87％的志愿者身份。

（3）大数据分析结果可能泄露隐私

大数据分析是一把双刃剑：一方面，大数据分析可以发现知识，提高决策的质量和效率；另一方面，大数据分析也可能挖掘出用户的隐私，导致个人隐私泄露。例如在电子商务中，网站的推荐系统通过对顾客的购物信息（如购物时间、购物地点、具体所购物品和购物数量）进行分析，可以发现顾客的购物喜好，进而对顾客进行个性化推荐，提高顾客的购物效率和购物体验。但是推荐系统也可以通过购物信息挖掘出某一位顾客所购食品的特征，并进一步推断出该顾客是否罹患糖尿病。

9.2　大数据采集的风险

高质量的数据是大数据分析可靠的基础，数据质量对于大数据分析具有十分重要的影响。目前评价数据质量的优劣有六个参考指标：①完整性（completeness）——度量遗失的数据以及不可用的数据；②规范性（conformity）——度量未按统一格式存储的数

据；③一致性(consistency)——度量是否存在歧义的数据；④准确性(accuracy)——度量是否存在不正确或者过时的数据；⑤唯一性(uniqueness)——度量重复数据或者属性重复的数据；⑥关联性(integration)——度量缺失或未建立索引的关联数据。大数据的采集阶段主要关注其中的完整性和准确性。

大数据的4V特征导致了大数据固有的复杂性，这给采集到真实、完整、准确的数据带来极大的挑战。

1. 采集的数据不准确

大数据时代虽然各种海量数据不断涌现，但数据量大并不代表这些数据都是有价值的。如果不根据分析的目标有的放矢地采集数据，则可能出现采集的数据不准确，影响大数据分析的结果。国内专家曾列举了一个错误的数据采集例子。2013年雅安地震以后，社交媒体(如微博、微信、人人等)的相关数据量激增，这些网站在短时间内就积累了海量的数据，但却很难反映地震区域全部的问题，因为社交媒体中有关雅安地震的数据大部分来自成都等大型城市。这很容易理解，大城市人口密度高，智能手机更加普及，网络覆盖也更广。而那些相对偏僻的地震灾区，收集的数据则少得可怜。由于电力、通信系统瘫痪，真正受灾最严重的地区却几乎统计不到相关的数据。因此，对地震的相关数据进行分析的时候，采用上述社交媒体的数据就不准确。

2. 采集的数据不完整

前面我们讲过，大数据价值密度低，即数据量越大，里面真正有价值的东西所占的比例就会越少。由于大数据来源复杂，应用需求也千差万别，锁定并采集这些少量有价值的内容无异于“沙里淘金”，稍有不慎则很容易忽略，从而导致采集的数据不完整。

3. 采集的数据不真实

大数据来源复杂，很多数据由不同的机构或组织提供，采集数据的时候很难掌控这些数据是否真实可靠。2014年，国外社交媒体Facebook的一份报告显示，其网站有7 600万个“僵尸账号”。在另一个社交媒体Twitter上，很多明星的僵尸粉丝(社交媒体中的虚假粉丝)数量更是惊人。据统计，著名演员贾斯汀·比伯的粉丝中有31%是僵尸粉，而著名流行歌手史蒂芬妮的僵尸粉更是占了34%。这些“僵尸账户”在社交媒体上造成了虚假的繁荣，很可能会让广告商对于明星账户的商业价值产生错误的评估，引起运营上的偏差，甚至导致投资上的失策。在电子商务领域，电子商务网站通过分析海量的顾客评论信息来向用户提供推荐服务，如在顾客选择商品时列出“最受好评的商品”“评分最高的商户”等内容。近年来，一些不良商家雇佣水军通过虚假的“刷好评”的方式来提高自己及商品的声誉，如果电子商务网站不加验证地利用这些虚假的好评，则会误导顾客消费，在经营上陷入危机。

因此，在采集大数据的时候，就需要从应用目标出发，明确以下问题：需要什么样的数据、这些数据是否足够、数据是从哪里来的、其中有多少数据是真正有价值的、这些数据有没有可能存在虚假信息等。从数据处理的第一个环节就开始减少误差对数据分析的干扰。

9.3　大数据处理与集成的风险

大数据的处理与集成主要是对已经采集到的数据进行适当的处理，清洗去噪以及进一步集成存储，将这些结构复杂的数据转换为单一的或是便于处理的数据结构，为以后的数据分析打下良好的基础。

大数据的处理与集成要为后续的数据分析提供高品质的数据，因此这个阶段输出的数据要尽可能满足数据质量的六个评价指标，即完整性、规范性、一致性、准确性、唯一性和关联性。从前面的内容中，我们知道大数据来源广泛，从各种渠道采集获取的数据不仅种类繁多、结构复杂，而且数据之中还存在歧义、冗余甚至错误，这给大数据的处理与集成带来一定的风险。

首先，数据的清洗去噪的尺度不容易拿捏。大数据时代数据具有价值密度低的特点，也就是说，大数据量并不意味着大信息量，很多时候它意味着冗余数据的增多和垃圾数据的泛滥。因此，对数据进行清洗和去噪是十分必要的，否则一方面过多的干扰信息会占据大量的存储空间，造成存储资源的浪费，另一方面这些垃圾数据会对真正有用的信息造成干扰，影响数据分析结果。大数据时代的数据清洗过程必须更加细致和专业，即在数据清洗过程中，清洗的粒度既不能过细，因为这会增加数据清洗的复杂度，甚至有可能会把有用的信息过滤掉（可能破坏数据的完整性或准确性）；清洗的粒度也不能过粗（可能导致数据冗余或者存在错误）。所以，在清洗过程中，清洗的尺度把握不好也会影响数据分析的质量。

其次，大数据的数据类型包含了结构化数据以及越来越多的半结构化数据和非结构化数据，目前还没有一项成熟的技术能够自动发现不同类型的数据之间的歧义或逻辑错误。例如，一个文本数据和一个图像内容之间是否存在歧义，或者视频内容与音频内容之间是否存在逻辑错误，都不能有效地检测出来，因此也会影响后续数据分析的质量。

最后，数据转换质量难以掌控。在大数据时代，数据呈现广泛的异构性，主要表现在以下几个方面。

（1）数据类型由传统的结构化数据为主逐步转向结构化、半结构化、非结构化数据三者的并存，而且半结构化、非结构化数据增长迅猛，所占的比重快速提高。

（2）数据的来源也逐渐多样化。传统电子数据的主要来源是机关、企业和学校等的服务器或者是个人电脑，这些设备位置相对固定，而且动态变化数据的比例不大。在大数据时代，随着互联网和移动设备在全球的普及以及物联网的应用，平板电脑、手机、各种传感设备等产生的数据爆炸式增长，而且这些数据随着时空变化而动态变化。

（3）传统的数据存储方式主要依靠关系型数据库，但这已经不足以满足大数据时代的数据存储需求。为了应对越来越多的海量数据和日渐复杂的数据结构，很多公司都开始研发适用于大数据时代的分布式文件系统和分布式并行数据库，如 Hadoop 的 HDFS、谷歌的 BigTable 等。在数据分析之前，数据格式的转换是必要的，要对这些动态变化并且具有广泛差异性的数据进行转换，过程是非常复杂和难以管理的。

9.4 大数据分析的风险

大数据处理和分析的终极目标是借助对数据的理解辅助人们在各类应用中作出合理的决策，这依赖于高精度的知识发现技术来对人类难以理解的底层数据特征进行深度挖掘和分析。由于大数据分析方法还处于“成长期”，还有很多不完善的地方，因此难以满足各种应用需求。

首先，传统意义上的数据分析主要针对结构化数据展开，且已经形成了一整套行之有效的分析体系。例如，利用数据库来存储结构化数据，在此基础上通过聚类、关联分析等方法构建数据分析模型来挖掘数据中隐含的知识。在面对大数据分析时，一方面由于半结构化和非结构化数据的存在，数据很难以类似结构化数据的方式准确构建出其内部的关系；另一方面海量数据流源源不断地到来，需要实时处理的数据很难有足够的时间去建立先验知识。

其次，目前的大数据分析方法不能胜任实时性的数据分析。在大数据时代，随着时间的流逝，数据中所蕴含的知识价值随之递减，实时处理成为大数据分析的典型需求。但是目前仍未存在一个通用的大数据实时处理框架，而且各种工具实现实时处理的方法各不相同，支持的应用类型也相对有限，这导致实际应用中往往需要根据自己的业务需求和应用场景对现有的技术和工具进行改造才能满足要求。

再次，数据融合存在困难。大数据时代数据来源多种多样，既有商业交易数据和科学研究数据，又有海量的社交媒体数据以及各种传感器产生的数据。每一种数据来源都有一定的局限性和片面性，只有对各种来源的原始数据进行融合才能反映事物的全貌，事物的本质和规律往往隐藏在各种原始数据的相互关联之中。数据分析时往往需要将这些不同来源的碎片化的数据进行融合，才能获得反映事物全貌的完整数据，这虽然可以增加数据挖掘的深度，但是目前还没有一个很好的技术能有效地将这些“一盘散沙”的数据充分整合，因为这些数据的格式千差万别，这就给数据融合带来相当大的困难。

最后，目前大数据分析技术还不能有效挖掘出隐藏在数据中的深层次知识。例如，目前大数据分析只能告诉我们用户正在做什么，而不能告诉我们他们在做的时候是怎么想的、背景是怎样的，或者有着什么样的情绪。例如，在社交网络中，通过大数据分析可以比较容易得到如下分析结果：一段时间内，某个用户和其他 3 个人每天对话超过 10 次，同时又和另外 10 个人经常发生互动，但是根据上述结果却很难分辨出这些人中间究竟哪些联系是真挚的情感与友谊的体现，而哪些联系只是为了应酬和生意。很多时候数字信息虽然比较严谨，但在做大数据分析的时候，更重要的是要挖掘出各种数字背后隐藏的各种深层次的知识，如语义、联系和情感等。

9.5 大数据解释的风险

对大数据分析结果的解释是大数据分析流程的最后一个环节，也是至关重要的环节，关系到用户能否进行正确的决策。如果数据分析的结果不能得到正确全面的解释

和理解,则会给数据用户造成困扰,甚至会误导用户。目前在大数据的解释技术特别是可视化展现技术方面还存在很多瓶颈问题和技术挑战,这客观上造成了大数据解释的风险。

首先海量异构的数据往往会有多种特征,在对分析结果进行解释的时候,关于这些特征的多重指标可能会导致分析结论的分歧。一组结果在不同人看来,会得出不同的结论。当一组结果反映很多关键指标的时候,这些多重指标会让分析者产生困惑,甚至得出与事实完全相反的结论。

其次,大数据时代的数据量大,分析更复杂,可视化技术是最佳的结果展示方式之一。当大数据以直观的可视化的形式展示在分析者面前时,分析者往往能够一眼洞悉数据背后隐藏的信息并转化成知识及智慧。但是可视化展示的效果除了跟数据有关以外,还与展示形式、人类视觉的敏锐性、分析者面对展示界面时的推断能力和信息搜索能力等因素都有关,其中任何一个因素都可能影响最终可视化分析的效果。另外,高维度的大数据通常需要降维处理才能以平面或者立体图形展示出来,但目前对于大规模、高维度和动态变化的数据,通过可视化技术动态来实时和精确地展示出来还是一个巨大的挑战。

最后,对大数据的解释除了采用可视化技术外,还强调交互式分析。计算机不仅要把大数据分析的结果以图形、图像的方式呈现给读者,还要能够和用户进行交互,根据用户的需求动态调整可视化的形式和内容,以满足用户的个性化需求。在大规模、高维度和动态变化的数据环境中,人机交互的界面如何动态调整是一个挑战。交互式分析还需要计算机精确理解用户对数据的需求,这又是另外一个挑战。

9.6　大数据的隐私和安全风险及其对策

目前,大数据是 IT 领域的研究和应用热点,受到了世界各国政府、学术界及工业界等社会各界的广泛关注,发展势头迅猛。但是大数据在提高社会和经济效益的同时,也给个人和组织的隐私以及数据安全带来极大的风险和挑战。例如,人们日常生活中的移动轨迹通常蕴含了个人的一些隐私信息(如家庭住址、工作单位、日常活动情况等)。在大数据环境下,掌握了这些移动轨迹数据,就能够很容易地分析发现这些隐私信息,从而导致个人隐私泄露。国外曾有研究表明,在 150 万条匿名的个人移动轨迹数据中,在不依赖外部其他背景知识的情况下,随机给出 2 个时空数据点,可以甄别出 50%的个人敏感轨迹(如他是否去过医院或者警察局),而如果给出了 4 个时空数据点,则被甄别出的敏感轨迹数据竟达到 95%。又如医学领域的基因研究中,通常需要收集和共享病人或者志愿者的基因数据。这些数据可以帮助医治心脏病、糖尿病等疾病,但是不可避免会涉及个人隐私。例如,通过 DNA 序列分析,可以推断出某人是否癌症患者。据麦肯锡公司的分析,如果把教育、交通、商业、金融、医疗卫生、石油、电力七个行业的数据公开用于大数据分析的话,可以带来 3 万亿美元的经济利益,但同时也会给相关个人和组织带来严峻的隐私泄露风险。因此,如何在充分利用大数据的同时不泄露用户的隐私,是一个非常重要的现实问题,关系到大数据的发展和应用。

大数据隐私是指个人或组织机构等实体不愿意被外部知道的敏感信息，包括个人的行为模式、兴趣爱好、位置信息、健康状况、财务状况等。大数据的隐私问题本质上来自大数据中的敏感信息的泄露，因此，保护大数据隐私最根本的目的就是保护敏感数据不被泄露。目前，国内外专家从如下两个角度来研究和解决大数据的隐私风险：大数据处理流程存在的隐私风险和大数据处理平台带来的隐私风险。

9.6.1 大数据处理流程的隐私风险

我们在本书第2章中介绍过大数据的处理流程包含数据采集、数据处理与集成、数据分析和数据解释四个阶段。数据采集负责各种数据源收集和存储所需要的信息；数据的处理与集成主要是对已经采集到的数据进行适当处理，正如前面一章所介绍的，包括消除冗余，清除不一致的数据以及进一步的集成存储；数据分析从大数据中挖掘发现有价值的模型或规则；数据解释主要通过可视化、数据溯源等技术来展示大数据的分析结果。在大数据的处理流程中，隐私风险主要集中在前面三个阶段。

1. 数据采集阶段的隐私风险

在大数据环境下，有许多个人数据也许是在用户不知情或未经同意的情况下被收集的。例如，一些商家在提供服务的同时也收集个人的购物记录、手机通话记录、个人移动轨迹、网站访问和登录记录等。2011年4月《纽约时报》报道，苹果公司通过iPhone系统在用户毫无觉察的情况下跟踪并收集用户的地理位置信息。另外，谷歌公司也通过浏览器在用户不知情的情况下收集用户的上网搜索记录，从而掌握用户的上网行为、政治倾向和消费习惯等。这些个人数据一方面可以帮助商家了解顾客的使用情况，从而更好地提供服务；但是另一方面，如果这些数据被不可信的商家收集或者贩卖给恶意的攻击者，则可能导致个人隐私泄露。例如，商家通过签到服务采集用户签到的地理位置信息，如果这些信息被非法恶意使用，则可能通过签到的位置信息以及签到的位置序列推测出用户的家庭住址、单位位置和移动轨迹等个人隐私数据。

上述在用户不知情或者未经同意情况下收集数据的现象在大数据时代非常普遍，隐藏着巨大的隐私风险。目前这类风险还缺乏法律法规的监管，主要依靠商家的自律和自觉遵守某些规范来确保用户隐私不被泄露。确保用户在其个人数据被采集时有知情权和授权允许，让用户能够随时掌握个人数据的使用情况，以及发现恶意使用后，用户如何及时销毁个人数据等，这些权利的实施还需要政府出台相关的法律法规。

2. 数据处理与集成阶段的隐私风险

数据处理与集成阶段的一个重要的工作，是把从各个分散的数据源采集到的数据进行集成和融合，从而更好地服务于数据的分析与管理。例如，商品零售商集成线上、线下的销售记录，可以获得消费者更多的信息，预测消费者的购物偏好；又如，地图导航服务提供商集成不同路段上的位置传感器数据，可以获得更好的道路规划和交通路线。然而多个数据源的集成与融合容易推断出用户的个人敏感信息，从而给隐私保护带来严峻的挑战。例如，图9-1所示是两个来源不同的数据记录，上面一条是病人的医疗记录，包含了用户名、身份证号、疾病、治疗方案、出生年月、性别和邮编，为了保护用户的隐私，其中用户名和身份证号用随机数字代替；下面一条是选民记录，包含了用户名、地址、出生年月、

性别和邮编。如果仅仅只有医疗记录，很难从中识别出一名具体的病人。但是如果同时拥有这两条记录并把它们相关联，如图中阴影部分内容所示，则可以通过选民记录里面的出生年月、性别和邮编与医疗记录里面的出生年月、性别和邮编进行匹配，推测出医疗记录里面的某一名具体的病人。

另外据报道，美国最大的互联网服务提供商之一美国在线服务公司(AOL)，为了保护用户的隐私，曾把用户搜索记录里面的名字和身份证号全部替换成随机数。照理说经过这种匿名和模糊化处理后，用户的隐私应该安然无恙，然而《纽约时报》的一名记者还是通过其他背景知识推断出其中一名用户是佐治亚州的一名寡妇。

医疗记录数据：

用户名（已匿名）	身份证号（已匿名）	疾病	治疗方案	出生年月	性别	邮编

选民记录数据：

用户名	地址	出生年月	性别	邮编

图 9-1　两个数据源的集成与融合导致的隐私泄露

3. 数据分析阶段的隐私风险

大数据的计算分析能力能够在海量数据中“大海捞针”，发现其中隐含的深层次的信息，导致隐私信息的泄露。例如，通过对用户移动轨迹的分析，可以挖掘出用户频繁发生的行为、行为之间的相关性以及用户行为的历史轨迹等，这不仅会泄露用户历史行为的隐私而且能预测用户未来的行为。又如，大数据下的电子商务网站可以利用其个性化推荐系统挖掘出用户的兴趣特点和购买行为，向用户推荐其感兴趣的商品和信息。然而，用户购买的商品信息和行为模式也很容易被电子商务网站挖掘出来，进而导致隐私信息的泄露。

大数据分析带来的隐私风险包括直接风险和间接风险。直接风险是指由数据分析结果可能泄露隐私信息，上述例子中的用户移动轨迹数据和购买行为的挖掘结果能够泄露隐私就属于这类；间接风险是指大数据分析方法可能导致原有的隐私保护方法的失败。例如，医疗数据库原来可以通过匿名或者模糊的方法来保护病人的隐私，但大数据的分析方法通过数据之间的关联，可以定位或发现具体某一位病人的信息，从而导致这些原有的隐私保护方法失效。

现在专家普遍认为，大数据分析的隐私风险主要来自三个方面。第一个是新型计算平台的强大处理能力。在大数据环境下，以 Hadoop 和 spark 为代表的计算框架具有强大的处理能力，能够以批处理或者流式处理方式并行处理海量的数据。第二个是基于这些计算框架开发出了更加快速的算法。例如，基于 Hadoop 的快速聚类方法 k-center 和 k-median、多维聚类方法 BoW 和关联聚类方法 Co-Cluster 等高性能的算法一方面能够深入分析大数据中细小的、彼此之间毫不关联的数据碎片，从而发现更深层次的知识；另一方面也为恶意分析者提供了发现隐私数据的快速方法。第三个是复杂的数据分析模型。以前单一的分类、聚类等模型已经不能应对大数据的海量数据和多样性，进而出现了更为

复杂和高效的数据分析模型,如基于随机优化的分类方法 SDCA 和回归分析方法 SAG。这些数据分析模型有助于从大数据中挖掘用户隐私。

9.6.2 大数据处理平台带来的安全和隐私风险

正如本书第 2 章所述,云计算技术是目前大数据存储和处理的重要平台(简称云平台,提供云计算服务的机构称为云服务提供商)。人们一方面可以利用云平台的存储能力来保存海量的大数据,另一方面又可以基于云平台的强大计算能力来分析和处理大数据。云平台虽然给大数据的应用和发展提供了强有力的支持,但同时也给大数据的应用带来了一定的隐私风险,主要体现在大数据的存储、搜索和计算三个方面。云平台隐私风险的最根本原因在于用户数据保存在云服务提供商完全掌控的云平台中,用户丧失了对数据的绝对控制权,而云服务提供商并不是完全可信任的。

1. 大数据存储面临的隐私风险

随着大数据的大量涌现,人们对存储空间的需求越来越大,在这种趋势下,基于云平台的存储方式也应运而生。这种存储方式利用云平台强大的存储能力,把数据存放到云平台中,使用者可以在任何时间、任何地方,通过任何可联网的装置连接到云上方便地存取数据。

基于云平台的大数据存储中,大数据的拥有者把自己的数据存储在云平台上后,云服务提供商或者非法入侵的黑客可以偷窥数据内容,还可以未经数据拥有者的同意把数据泄露给其他未授权的第三方,从而导致隐私数据的泄露。近年来,由于黑客的非法入侵和云平台管理员的不当操作造成了多起云安全事故,直接导致了大量用户资料和私人数据的泄露。例如,谷歌公司在 2011 年由于黑客入侵,发生了 Gmail 大规模用户数据泄露事件。另外,由于用户丧失了对数据的绝对控制权,云服务提供商可以非法修改、删除或添加数据的内容,从而破坏数据的真实性和完整性。为了避免这些隐私风险,通常采用加密的方法来确保数据隐私不被泄露并防止数据被非法修改和破坏,具体方法将在下一节介绍。

2. 大数据搜索面临的隐私风险

大数据的拥有者把自己的海量数据存储在云平台上后,为了高效管理和利用数据,需要对这些数据进行搜索。如上所述,为了避免数据隐私泄露,这些数据都是加密后以密文的形式保存在云平台上的。要在这些加密的数据上完成检索工作,数据的使用者有两个选择。第一个选择是把云平台上保存的加密数据全部取回本地,解密后再用关键词检索。这种方法效率非常低,不仅下载过程会占用过多的网络带宽,也会占用过多的本地存储,而且解密过程还会消耗大量的本地计算资源;另一种极端的方法是把关键词和加密数据的密钥提供给云平台,让云平台把数据解密,然后在明文数据上根据关键词检索,这无疑又会让原来加密保护的数据重新曝光在云服务提供商或者非法用户的视线之下,从而泄露数据的隐私。

3. 大数据计算面临的隐私风险

云计算的强大处理能力是大数据发展和应用的重要支撑,在大数据环境下,数据拥有者或者其他用户通常希望利用云平台强大的计算能力分析处理大数据并将计算结果返

回，然而作为计算输入的大数据或者计算结果可能是非常机密的，如果不加以保护，则云服务提供商能够知晓这些数据，从而导致隐私泄露。

9.6.3　保护大数据隐私和安全的对策

由于大数据及其处理流程的特点，传统的隐私保护理论和技术已不能很好地避免大数据处理流程中的隐私泄露。人们已经开发了一些行之有效的隐私保护技术，下面我们选择其中重要的技术进行介绍。需要说明的是，目前没有一种万能的方法能够解决所有隐私问题，每一种方法都有自己的优缺点和应用场景。

1. 匿名化技术

匿名化是保护隐私的重要技术，它的思想是通过隐藏或者模糊的方法使数据不能被精确识别，例如生活中可以用随机数或者其他字符替换人的名字来实现匿名。K-匿名技术是一种重要的匿名技术，它的思想可以通过在基于位置的服务中的应用示例来说明。如图 9-2 所示，假设 K 是 5。在基于位置的服务中，如果一个用户想要查询某一个医院的坐标，位置服务器收到用户的查询请求后，并不是只返回该医院的精确坐标给用户，因为观察到这个返回数据的人会推测出用户可能去医院看病，从而泄露用户的隐私。相反，位置服务器返回给用户的位置信息是包含了该医院在内的一个区域内（如图 9-2 中的圆形区域）的 5 个地理位置的坐标（如图 9-2 中的其他黑点），再让用户从中找到自己需要的位置。其他人只知道这 5 个位置信息，但是不知道用户具体需要的是哪一个位置，从而起到了保护用户隐私的目的。

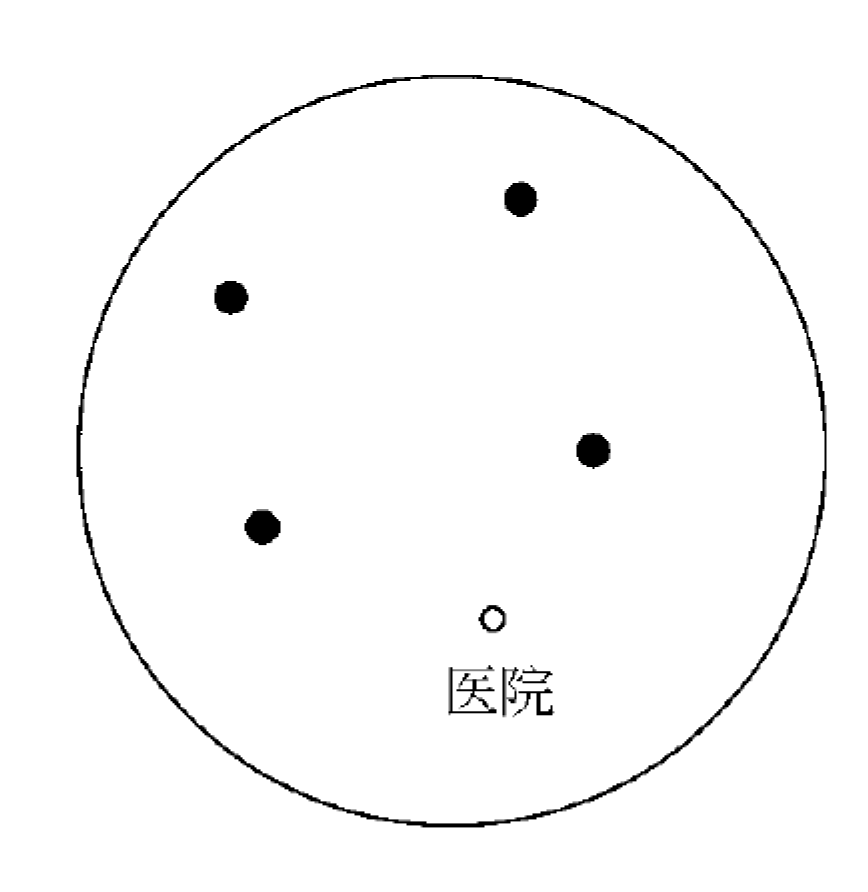


图 9-2　K-匿名技术示例

2. 数据加密存储

当大数据保存在云平台中时，为了保护数据的隐私不被泄露，采用加密技术来确保除了数据拥有者和授权用户能够访问数据明文以外，包括云服务提供商在内的其他人都无法得到数据明文。数据加密存储和访问授权如图 9-3 所示，图中包括三个参与者，分别是数据拥有者、云服务提供商和用户。数据加密存储和访问授权的过程如下：

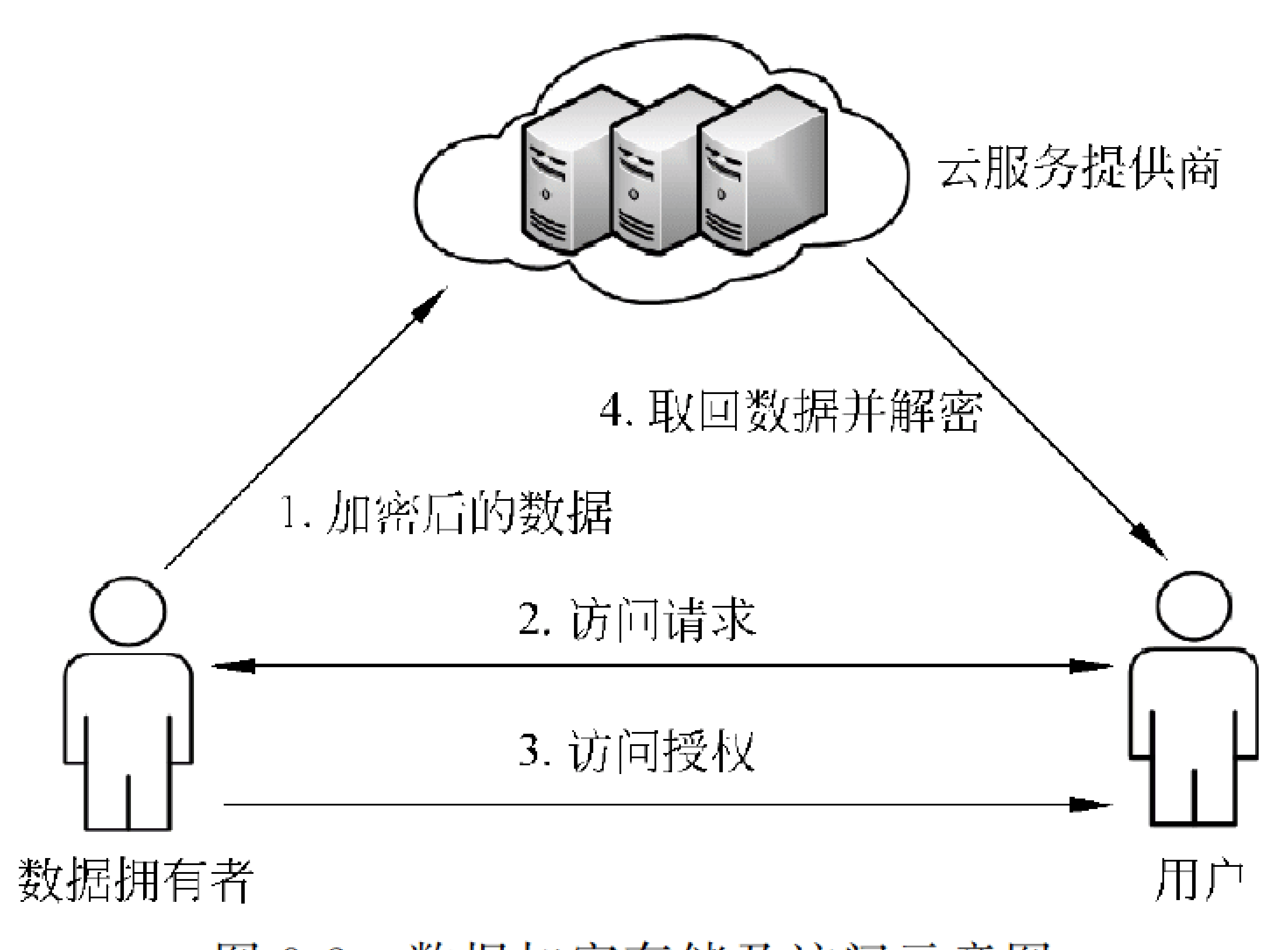


图 9-3　数据加密存储及访问示意图

(1) 数据拥有者把数据加密后上传到云平台保存,由于不知道加密的密钥,云服务提供商以及非授权的用户都不知道数据明文,所以数据的隐私得到保护;

(2) 当某位用户需要访问这些数据时,他先向数据拥有者发起访问请求;

(3) 数据拥有者如果授权该用户访问这些数据,则把数据加密的密钥发送给用户;

(4) 用户从云平台取回数据后,用获得的密钥解密数据,获得数据明文。

3. 数据完整性保护

数据完整性保护是指数据拥有者把数据保存在云平台上后,数据不能有丝毫遗失或损坏,也不能被伪造或者篡改。数据完整性保护的思想如图 9-4 所示。

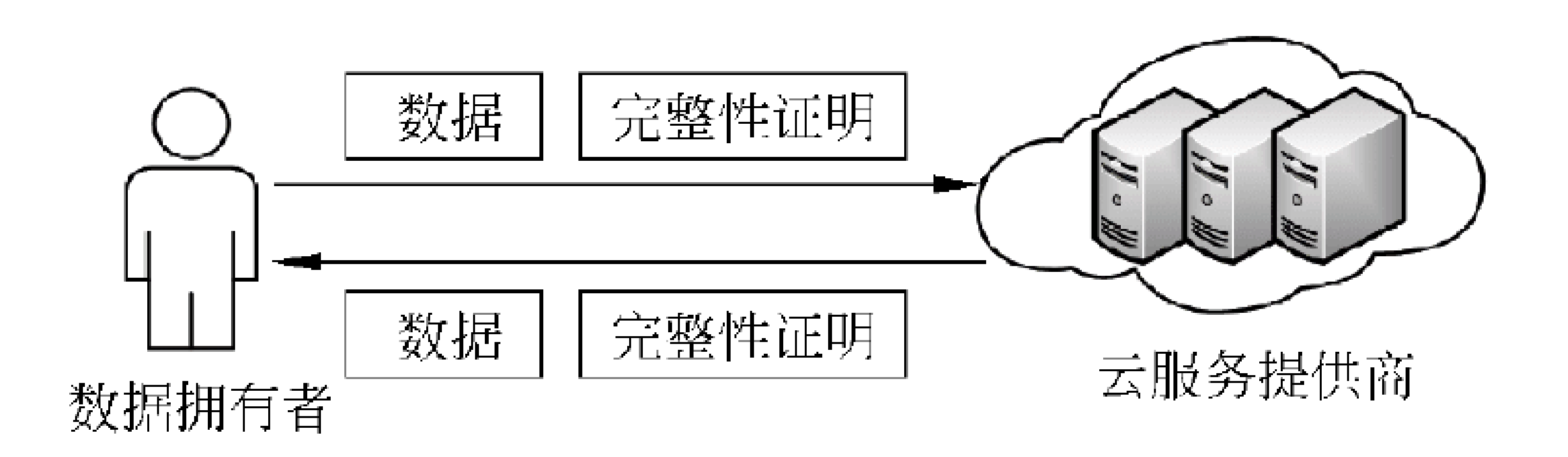


图 9-4 数据完整性保护示意图

首先,数据拥有者在上传数据到云平台之前,根据数据内容采用密码技术生成一个不可伪造或篡改的完整性证明,然后把数据(或者加密后的数据)和相应的完整性证明一起上传到云平台中保存。当数据拥有者需要查验保存在云平台上的数据是否存在遗失或者伪造等情形时,他从云平台取回数据和相应的完整性证明,再根据取回的数据重新计算一个完整性证明,把这个新计算的完整性证明和原来保存在云平台上的完整性证明进行比对,如果不一致,就说明数据存在遗失或者篡改。

4. 同态加密技术

同态加密技术是一种新型的加密技术,它的特点是直接对加密数据进行诸如计算、比较等操作,得出正确的结果,而在整个处理过程中无须对数据进行解密,输入的数据和输出的处理结果全部以密文的形式存在,只有拥有密钥的用户才能解密获得处理结果。同态加密既保证了输入数据的安全,又确保处理结果的隐私不被泄露,因此特别适合基于云平台的大数据处理的隐私保护。同态加密的原理如图 9-5 所示,图中数据拥有者希望利用云平台强大的计算能力计算两个数 X1 和 X2 相加的和,但是不希望云服务提供商知道 X1 和 X2 以及运算结果的具体内容。数据拥有者可以采用同态加密方法实现这一目的。他先分别加密 X1 和 X2 获得相应的密文 C1 和 C2,然后把 C1 和 C2 上传到云平台,云平台基于 C1 和 C2 通过同态加密技术计算得到 X1 和 X2 相加的和的密文 Cs。数据拥有者从云平台获得 Cs 后,解密 Cs 得到 X1 和 X2 相加的结果。整个处理过程中云平台无须对 C1 和 C2 进行解密,计算结果也以密文形式存在,只有拥有密钥的数据拥有者才能解密

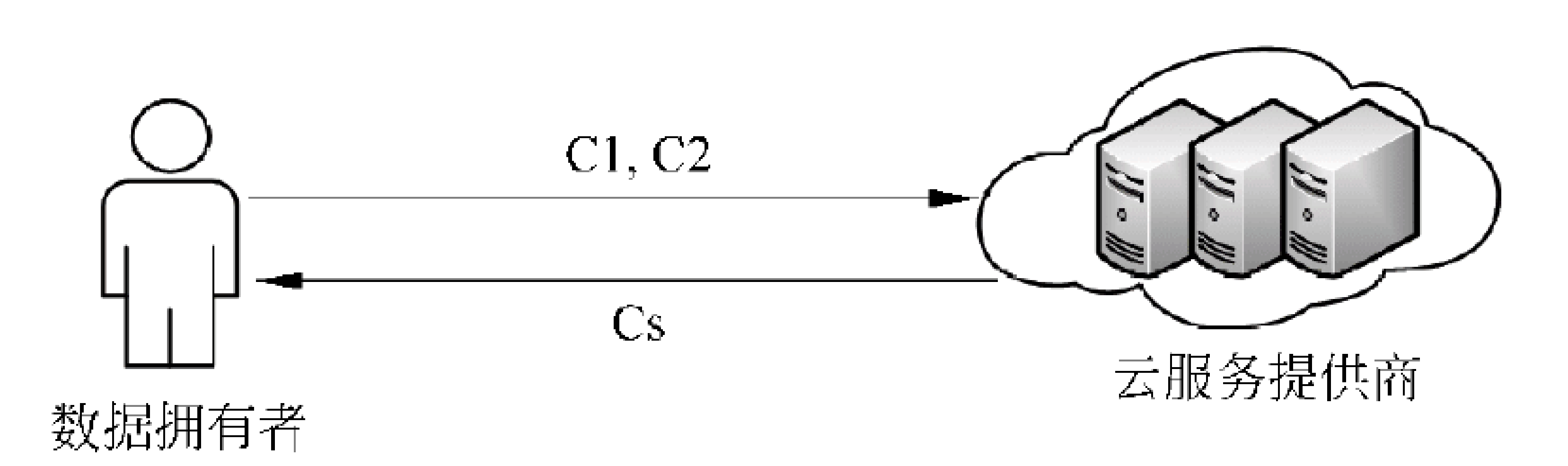


图 9-5 同态加密技术示例

获得处理结果。

5. 保护隐私的信息检索技术

如前所述，为了保护大数据的隐私不被泄露，通常把数据加密后再存储在云平台中。从用户的角度来看，接下来一个重要的工作就是在这些加密的数据中进行检索，以方便使用和维护这些大数据。保护隐私的信息检索技术的目的是让用户从保存在云平台上的加密数据中检索需要的数据，并且不会泄露数据以及检索关键词的内容。这里值得一提的是检索用的关键词的内容也要防止隐私泄露，如果关键词的内容泄露了，那么云服务提供商或者其他非授权的人员可以根据关键词的内容，推测出检索出来的加密文件的内容。保护隐私的信息检索技术的原理如图 9-6 所示，分成如下四个步骤。

（1）数据拥有者用密钥加密数据及其索引词，然后把加密后的数据以及加密后的索引词上传到云平台保存，由于数据及其索引词都是加密的，所以它们的内容不会泄露给非授权的人（如云服务提供商）；

（2）当数据拥有者需要检索某一个数据的时候，他先采用加密算法加密检索的关键词，并把加密后的关键词上传到云平台服务器进行检索；

（3）云平台收到加密的关键词后，结合保存在云平台中的加密的索引词，找到相应的数据并把它下传给数据拥有者；

（4）数据拥有者解密收到的数据，就获得了所需要的检索内容。

检索过程中关键词、索引词和检索到的数据都是加密的，所以云服务提供商及其他非授权访问的人都不知道检索的内容。从而保护了数据的隐私。

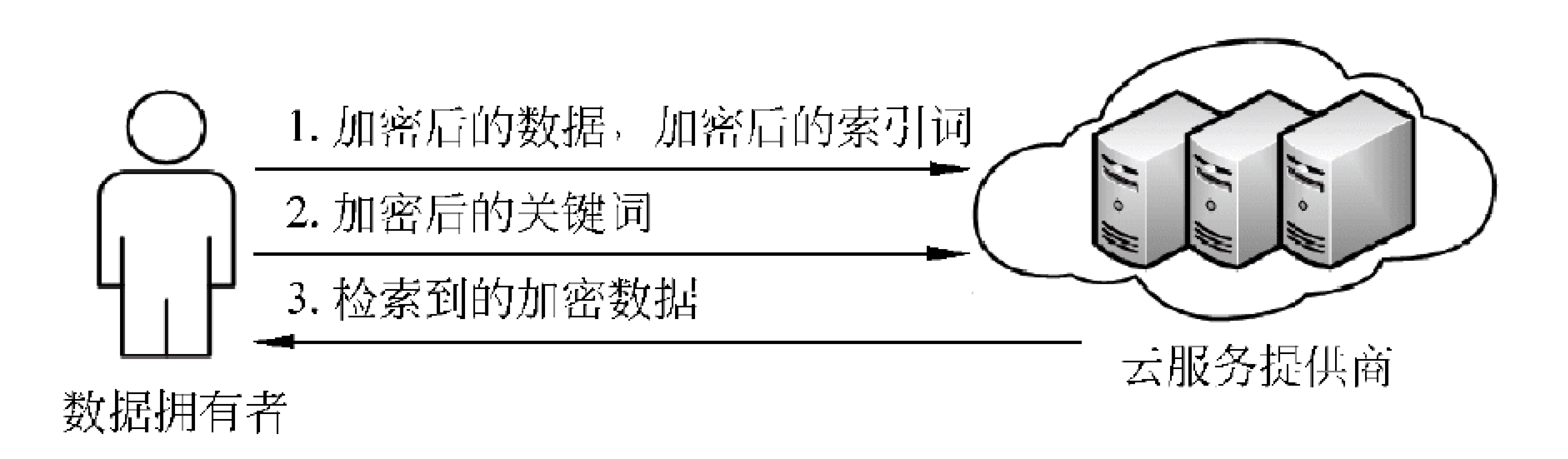


图 9-6　保护隐私的信息检索技术示意图

参考文献

[1] 孟小峰，张啸剑. 大数据隐私管理[J]. 计算机研究与发展，2015，52(2)：265-281.

[2] 黄刘生，田苗苗，黄河. 大数据隐私保护密码技术研究综述[J]. 软件学报，2015，26(4)：945-959.

[3] 谭霜，贾焰，韩伟红. 云存储中的数据完整性证明研究及进展[J]. 计算机学报，2015，38(1)：164-177.

[4] 曹珍富，董晓蕾，周俊，沈佳辰，宁建廷，巩俊卿. 大数据安全与隐私保护研究进展[J]. 计算机研究与发展，2016，53(10)：2137-2151.

[5] 冯登国，张敏，李昊. 大数据安全与隐私保护[J]. 计算机学报，2014，37(1)：246-258.

[6] 傅颖勋，罗圣美，舒继武. 安全云存储系统与关键技术综述[J]. 计算机研究与发展，2013，50(1)：136-145.

[7] 李经纬，贾春福，刘哲理，李进，李敏. 可搜索加密技术研究综述[J]. 软件学报，2015，26（1）：109-128.
[8] 马瑕. 对大数据分析相关问题的思考[J]. 信息通信技术，2013，6：58-62.
[9] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，(1)：146-169.
[10] 刘智慧，张泉灵. 大数据技术研究综述[J]. 浙江大学学报：工学版，2014，(6)：957-972.

第10章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过9个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低170千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是NASA的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是NASA指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现170千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致NASA 3.28亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章 大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1 大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位"磅"，而不是 NASA 指定的"牛顿"。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章 大数据治理简介

通过前面章节的学习,我们已经越来越清楚地看到,随着计算机技术及网络通信技术的普及和发展,人类进入了大数据时代。我们每天面临不断增加的海量数据,包括电子商务的网上交易数据,Web 网页的内容,社交媒体的文本、视频和图片,物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源,对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点,因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下,大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用,并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容,然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1 大数据治理的必要性

大数据治理(big data governance)是用好大数据资源,充分发挥大数据价值的重要手段。通俗地讲,大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合,具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一:美国火星气象卫星发射失败。发射太空探测器需要大量的数据,如果这些数据的可靠性没有保证,则会带来灾难性的后果。1999 年,美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行,在切换进入火星轨道之后,卫星突然意外地进入了比预定高度低 170 千米的火星轨道,最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落,并燃烧殆尽。事后经过调查,事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”,而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据,从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差,从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二:自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头,它的作用是探测汽车行驶过程中的环境变化,并及时作出反应。例如,摄像头发现了停车标志,就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理（big data governance）是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局（NASA）发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习,我们已经越来越清楚地看到,随着计算机技术及网络通信技术的普及和发展,人类进入了大数据时代。我们每天面临不断增加的海量数据,包括电子商务的网上交易数据,Web 网页的内容,社交媒体的文本、视频和图片,物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源,对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点,因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下,大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用,并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容,然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源,充分发挥大数据价值的重要手段。通俗地讲,大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合,具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一:美国火星气象卫星发射失败。发射太空探测器需要大量的数据,如果这些数据的可靠性没有保证,则会带来灾难性的后果。1999 年,美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行,在切换进入火星轨道之后,卫星突然意外地进入了比预定高度低 170 千米的火星轨道,最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落,并燃烧殆尽。事后经过调查,事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位"磅",而不是 NASA 指定的"牛顿"。卫星发射前工程师并没有检查数据,从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差,从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二:自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头,它的作用是探测汽车行驶过程中的环境变化,并及时作出反应。例如,摄像头发现了停车标志,就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

第 10 章　大数据治理简介

通过前面章节的学习，我们已经越来越清楚地看到，随着计算机技术及网络通信技术的普及和发展，人类进入了大数据时代。我们每天面临不断增加的海量数据，包括电子商务的网上交易数据，Web 网页的内容，社交媒体的文本、视频和图片，物联网和可穿戴设备产生的各种各样的传感数据等。数据已成为宝贵的资源，对于科学研究、企业经营和国家管理都具有重要的战略意义。由于大数据具有容量巨大、结构复杂多样、价值密度低的特点，因此如何管理好并充分利用这些海量数据成为大数据时代的迫切需要。在这种背景下，大数据治理应运而生。大数据治理的目的就是要保证数据资源的正确、可靠、安全、可用，并充分发挥大数据的价值。本章简要介绍大数据治理的概念和内容，然后通过一个数据质量控制的案例来说明大数据治理的思想。

10.1　大数据治理的必要性

大数据治理(big data governance)是用好大数据资源，充分发挥大数据价值的重要手段。通俗地讲，大数据治理是组织内部管理好和使用好大数据并使之成为战略资产的一个规范和政策的集合，具体内容包括维护和提高数据质量、数据资源的保值和增值、保护数据的安全和隐私、规范用户使用数据的行为和追责、协调组织内部使用数据资源的需求等。

下面我们通过两个案例来理解大数据治理的必要性。

案例一：美国火星气象卫星发射失败。发射太空探测器需要大量的数据，如果这些数据的可靠性没有保证，则会带来灾难性的后果。1999 年，美国国家航空航天局(NASA)发射了一颗火星探测气象卫星。经过 9 个月的飞行，在切换进入火星轨道之后，卫星突然意外地进入了比预定高度低 170 千米的火星轨道，最终这颗卫星因为不能承受火星低纬度大气的强烈摩擦而坠落，并燃烧殆尽。事后经过调查，事故的原因是 NASA 的工程师在设计卫星的时候使用的测量单位是英制单位“磅”，而不是 NASA 指定的“牛顿”。卫星发射前工程师并没有检查数据，从而未发现这个错误。这两个测量单位之间的误差最终使卫星的轨道高度计算出现 170 千米的巨大偏差，从而导致卫星发射失败。这个看似很小的错误导致 NASA 3.28 亿美元的损失并使美国的太空探索推迟了数年。

案例二：自动驾驶汽车识别交通标志出现错误。以深度学习为代表的新一代人工智能技术在自动驾驶汽车上获得了成功应用。自动驾驶汽车上有一个充当驾驶员眼睛的摄像头，它的作用是探测汽车行驶过程中的环境变化，并及时作出反应。例如，摄像头发现了停车标志，就应该及时触发制动使汽车停止运行。这需要利用大量不同的交通标志图

像数据和深度学习技术,建立一个交通标志识别系统,使摄像头能够自动识别交通标志,如图 10-1 所示。

据国外研究报道,目前在建立交通标志识别模型的时候,都没有判断交通标志图像数据真实与否,因此黑客能够通过一些刻意伪造的交通标志图,误导识别模型,导致出现判断错误。如图 10-2 所示,图中左边是一个正常的停止标志图像,右边是黑客刻意伪造的停止标志图像,二者人眼几乎无法分辨,但是黑客可以利用伪造的图像误导识别系统,使它把停止标志识别成其他交通标志,从而带来灾难性的后果。

从上述两个案例可以看出,在大数据时代,如果无法保证数据质量,包括正确、精确、真实、可靠和及时,不仅不能充分发挥大数据的价值,而且会带来严重的后果。因此,大数据治理势在必行。

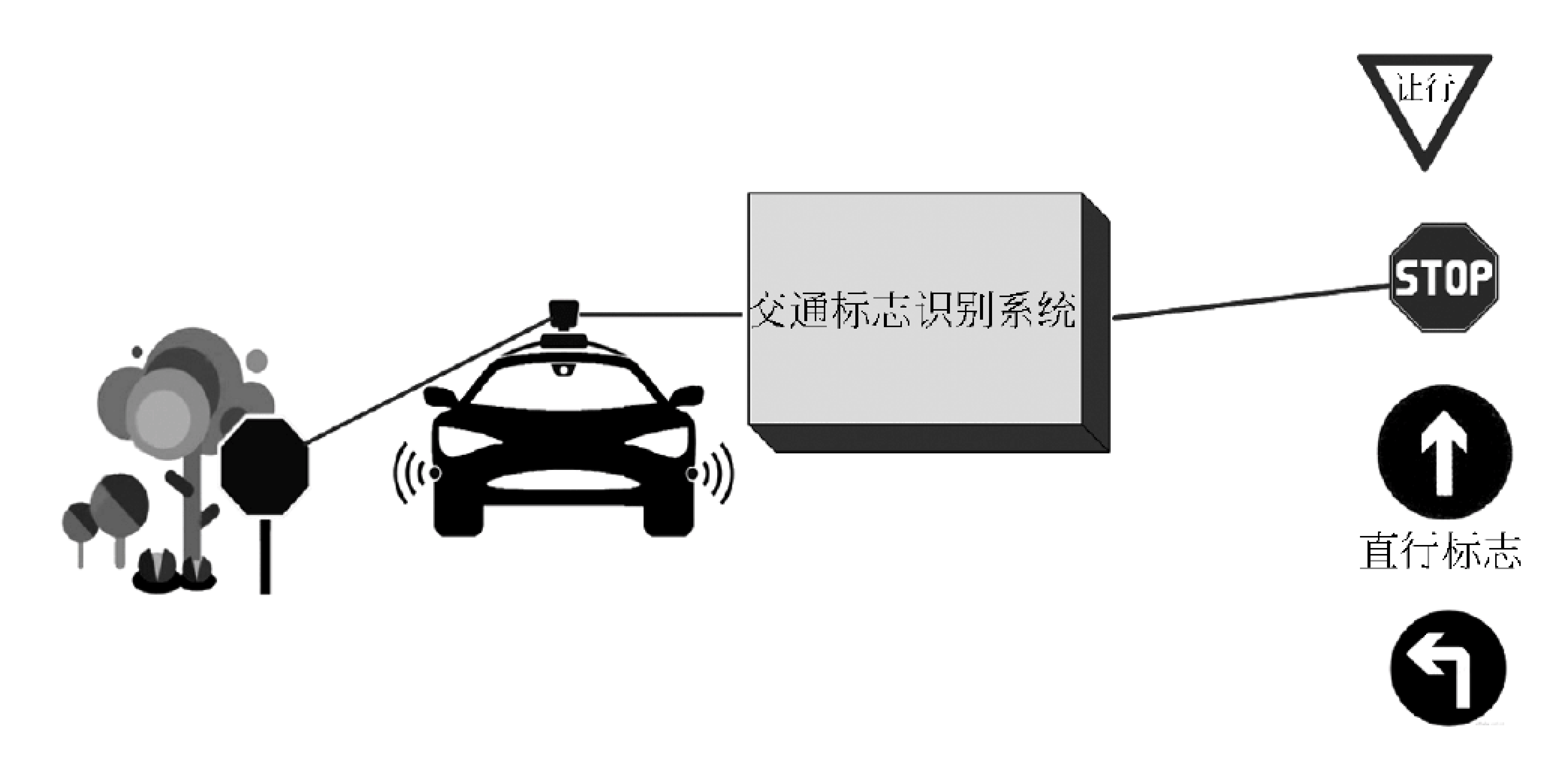


图 10-1　自动驾驶车辆交通标志识别示意图



图 10-2　正常和伪造的交通标志图像

10.2　大数据治理的概念

大数据治理的概念是从信息系统中的信息治理(information governance)的概念发展而来的。目前关于信息治理没有统一的定义,专家从不同的角度来阐述它的内涵。维基百科把信息治理定义为:一个多学科交叉的组织结构、政策、流程和控制措施的集合,用来管理组织内部的信息资源,以满足组织当前和未来在法律法规、风险控制、环境和操作方面的要求。著名的信息咨询公司加德纳公司把信息治理定义为:信息治理是关于信息的决策管理和责任追究的规范说明,包含了相关的过程、角色、标准和度量,用以规范在信

息的产生、存储、使用、存档、删除和评价中的有关行为，以帮助组织在实现有关目标的过程中高效和有效地使用信息。

目前，大数据治理还是一个崭新的领域，《大数据治理》一书的作者 Soares 给出一个比较权威的定义：大数据治理属于广义的信息治理范畴，它通过协调组织内部不同部门的目标，制定相关政策规范，以满足大数据优化、安全隐私保护和盈利的要求。Soares 的上述定义短小精悍，包含下面六个方面的含义。

1. 大数据治理属于广义的信息治理范畴

组织机构应该拓展当前信息治理的范畴，把大数据的内容也考虑在内，如聘用大数据的管理人员和分析人员，增加对大数据的元数据、主数据和安全隐私保护的管理等。

2. 大数据治理要制定相关的政策规范

这些政策规范要明确规定在不同环境下如何使用大数据，而且这些政策规范必须符合组织内部的法律法规要求。

3. 大数据的优化

大数据治理要帮助组织在如下方面优化和提高大数据的质量，包括：建立元数据，方便大数据的管理和使用；定期清理和维护大数据，确保数据质量；实施信息生命周期管理，及时清理不需要的数据信息。

4. 大数据的安全和隐私保护

大数据治理要帮助组织建立保护数据安全和隐私的政策措施，如保护数据不被非法访问，数据不会泄露隐私以及数据不会被非法修改和破坏等。

5. 大数据要能够盈利

大数据治理要帮助组织实现通过大数据来盈利，包括把数据出售给第三方盈利或者利用大数据开发新的增值服务。

6. 协调组织内部不同部门之间的目标

组织内部各个部门之间对如何使用数据的规定和要求可能不一样，甚至会出现冲突的情况。大数据治理要能够协调和统一这些规定和要求。

有的文献中经常出现如下三个容易混淆的术语：IT 治理(IT governance)、信息治理和大数据治理，这是三个相互联系又相互区别的概念。

IT 治理是公司治理的一个部分，关注组织内部信息系统的建设、开发、运行、性能和风险管理。它的目的是保证信息系统的运行满足高效、有效、安全和经济的目标。IT 治理的对象除了包括信息系统软件和硬件的采购、开发、运行和维护，还包括其中的信息管理。因此，IT 治理包含信息治理和大数据治理；信息治理是从 IT 治理分离出来的一个分支。随着组织内部信息的不断增加，信息的重要性日益凸显，于是出现了信息治理的有关政策、流程和方法来对这些信息进行管理和监控。信息治理专注于维护信息质量，提高信息使用的价值并降低使用信息的风险，信息治理包括元数据管理、主数据管理、数据质量管理、信息生命周期管理、隐私保护、风险管理等；大数据治理是信息治理在大数据时代的拓展，除了包含传统信息治理的以上内容，大数据治理还需要考虑大数据的一些特点，包括数据量大、结构千差万别、价值密度低、数据动态变化、数据来源广泛、相关性分析比重大等。另外，大数据治理还增加了一些特色内容，如使用大数据的行为规范和追责、大数

据优化、大数据的安全和隐私保护、大数据保值增值以及协调不同部门之间使用与管理大数据的目标和需求。

10.3 大数据治理的核心内容

在实施大数据治理的过程中，首先要明确什么是大数据治理所覆盖的内容，这样才能有的放矢地制定相关的政策和措施。目前大数据治理处于发展阶段，还没有正式的标准指南和细致入微的政策、方法和措施。《大数据治理》一书的作者 Soares 在其著作中给出了大数据治理的一个参考指南，提出大数据治理应覆盖的八个方面，如表 10-1 所示。

表 10-1 大数据治理应覆盖的八个方面

大数据治理应覆盖的八个方面	含 义
组织管理	大数据治理应该关注企业在原有的组织结构和工作职责方面是否把大数据纳入考虑范围，如是否配备专门的人员来承担相应的工作、是否有大数据的应急管理措施等
元数据管理	元数据是描述大数据的数据，主要用来描述大数据的特征属性。通过元数据可以对大数据资源进行有效的组织和管理，如查找和定位大数据资源、记录和追踪大数据在使用过程中的变化等。大数据治理应该考察企业内部的元数据的管理措施是否到位，如元数据是否完整、元数据是否及时更新、元数据之间是否存在歧义等
安全和隐私保护	大数据治理要考察企业内部对大数据安全和隐私保护的政策及措施是否到位，如是否分类识别出敏感数据、数据是否加密保护、是否防止非法访问、是否防止数据被非法修改和删除等
数据质量管理	大数据的价值在很大程度上取决于数据的质量。大数据治理要考察企业内部维护数据质量的政策和措施，包括数据质量的度量标准、维护数据质量的政策和方法、验证数据真实性和完整性的技术和措施等
主数据管理	主数据可以理解为企业内部为完成一个决策所采用的全部数据资源的统称。主数据可以是一个文件，也可以是分布在不同区域的不同类型的数据。主数据可以是关系数据库数据或文本，也可以是声音、图像和视频。大数据治理要关注如下内容：数据分析方法是否能获得准确的结果，分析过程是否高效，所采用的主数据在内容方面是否完整(特别对于分布在不同区域的数据)，在时间方面是否为最新数据，不同格式的数据如何转换成一致的数据格式；不同的数据之间如何消除冗余和歧义等
数据生命周期的管理	大数据治理应该关注数据生命周期的管理政策和措施，包括是否记录大数据的来源和流动过程、如何识别有价值的数据并加以应用、如何及时删除或存档不再需要的数据以减少维护成本等
大数据的盈利和增值	大数据治理要关注企业如何通过大数据来实现盈利的政策和措施，包括把数据出售给第三方以获取盈利或者利用大数据开发新的增值服务
协调不同目标和需求	大数据治理要关注当不同企业之间以及企业内部各个部门之间使用数据的规定和要求不一致或者出现冲突的时候，如何处理这种不一致和冲突的情况，包括是否存在隐私泄露、是否存在法律风险等

在实际使用过程中，还需要进一步细化上述八个方面的内容，明确具体的治理目标和

策略，这需要同时考虑三个方面的因素，分别是大数据应用的行业或部门、大数据的类型以及信息治理的核心内容。这构成了大数据治理的框架，如图 10-3 所示。它们的含义如下。

1. 大数据应用的行业或部门

大数据治理是与大数据应用的行业或部门相关的，不同行业或部门的大数据应用不一样，相应的大数据治理的策略和内容可能也不一样。例如，对于人类基因数据，管理基因数据库的部门可能更加关注基因数据的隐私保护，而使用基因信息来研制药物的机构可能更关心基因数据是否真实和完备。

2. 大数据的类型

大数据的类型分为 Web 和社交数据、传感器数据（如 RFID 或 GPS 数据）、生物特征数据（如指纹或 DNA）、交易数据（如电子商务交易记录或者银行消费记录）以及个人创建的数据（如电子邮件、办公文档、调查报告）。不同类型的大数据对治理的目标和策略不一样。

3. 信息治理的核心内容

信息治理的核心内容包括组织管理、元数据管理、安全和隐私保护、数据质量管理、主数据管理、信息生命周期的管理，它们的含义如表 10-1 所示。不难理解，针对不同行业或不同的大数据类型，信息治理的核心内容可能不一样。

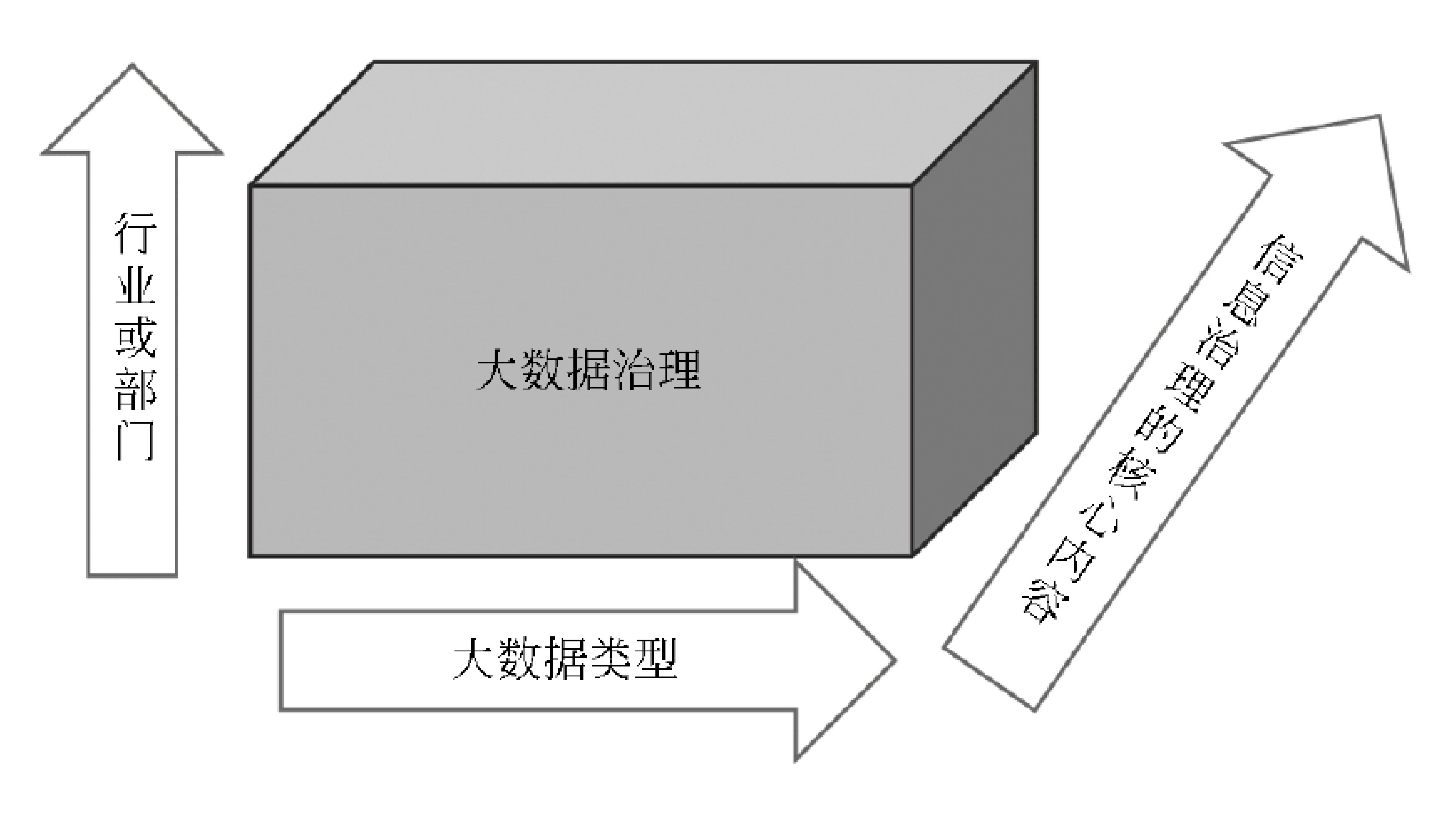


图 10-3　大数据治理的框架

10.4　案例

在大数据治理中，数据质量管理是一个非常重要的内容。数据质量对于大数据分析的结果至关重要，只有在数据是正确的、完整的、真实的前提下，大数据分析方法才能充分挖掘其中隐藏的知识，发挥大数据应有的价值，否则大数据分析的质量会下降甚至给出错误的分析结论。下面通过一个 IT 审计工作的案例来简要介绍数据采集过程的质量控制方法，虽然其中使用的数据是 Oracle 数据库系统中的结构化数据，但是其思想对大数据也是适用的。

10.4.1 工作思路

在审计工作中,审计人员需要从被审计单位采集数据。为了保证后续审计分析结果的准确性和客观性,审计人员必须对所采集数据的真实性和完整性进行验证。其中真实性验证是保证数据未经过被审计单位的人为改动,尽可能地判断出有无蓄意对数据进行增加、删除、篡改内容;完整性验证是保证所采集到的数据是完整的,被审计单位没有隐瞒或者屏蔽数据等情况。

在发出数据需求说明书后,一般应当在审计人员的监督下,由被审计单位技术人员完成实际的数据采集工作。在实施采集之前,审计人员应该首先对被采集数据的真实性和完整性的验证工作做好准备,然后再进行数据采集。

由于被审计单位使用的数据库管理系统不尽相同,验证数据真实性和完整性的方法也有所不同。在验证的过程中,审计项目组应该采用多种技术方法,从多个角度对数据进行验证。验证的角度可以是纯技术性的、纯业务性的或者技术与业务相结合的。数据验证可以在数据采集的同时或数据采集之后进行。从技术的角度出发,一般在数据采集的同时进行数据验证,下面将举例说明。一般而言,在进行了技术性验证后,往往需要从数据的经济含义出发进行验证,如经济总量、分量的核对,钩稽关系、借贷平衡等方面的验证。

在某些情况下,由于不能接触被审计单位的生产机系统,或由于数据是从备份介质中通过应用系统恢复,再导出为文本文件提供,难以在采集的同时直接对数据进行验证。在审计金融企业时,通常都会遇到这种情况。此时,就必须通过利用电子数据计算有关金额,与现有的纸质资料,如有关财务报表、日常统计表的相应金额比对,如有差异,则应进行分析,找出产生差异的原因。有时候金额有差异,并不表明数据一定不真实,还应进一步分析产生差异的原因。例如,如果获取了某银行明细账户的数据,则可以统计出年末总账余额,将统计出的数据与该行年末业务状况表的相应总账科目余额相比较,如通过电子明细账统计出的总账科目余额小于业务状况表的金额,不能立即得出数据不完整的结论,因为可能存在有部分业务未上机,还是手工处理的情况。

10.4.2 数据真实性的验证方法

1. 验证数据库的创建日期

由审计人员和被审计单位的技术人员一起,由审计人员监督、被审计单位技术人员执行,在被审计单位的 Oracle 数据库管理系统中,在 SQL* PLUS 工具中输入如下命令可以查看当前连接的数据库的创建日期。

```
Select Created From V$ Database;
```

结果如图 10-4 所示。

在得到了被审计单位的表空间后,审计人员可以进一步了解表空间的物理存储位置和所占的空间大小,这可以通过下述语句实现:

```
select  tablespace_name  表空间, file_name  物理文件名,
```

10.4.1 工作思路

在审计工作中，审计人员需要从被审计单位采集数据。为了保证后续审计分析结果的准确性和客观性，审计人员必须对所采集数据的真实性和完整性进行验证。其中真实性验证是保证数据未经过被审计单位的人为改动，尽可能地判断出有无蓄意对数据进行增加、删除、篡改内容；完整性验证是保证所采集到的数据是完整的，被审计单位没有隐瞒或者屏蔽数据等情况。

在发出数据需求说明书后，一般应当在审计人员的监督下，由被审计单位技术人员完成实际的数据采集工作。在实施采集之前，审计人员应该首先对被采集数据的真实性和完整性的验证工作做好准备，然后再进行数据采集。

由于被审计单位使用的数据库管理系统不尽相同，验证数据真实性和完整性的方法也有所不同。在验证的过程中，审计项目组应该采用多种技术方法，从多个角度对数据进行验证。验证的角度可以是纯技术性的、纯业务性的或者技术与业务相结合的。数据验证可以在数据采集的同时或数据采集之后进行。从技术的角度出发，一般在数据采集的同时进行数据验证，下面将举例说明。一般而言，在进行了技术性验证后，往往需要从数据的经济含义出发进行验证，如经济总量、分量的核对，钩稽关系、借贷平衡等方面的验证。

在某些情况下，由于不能接触被审计单位的生产机系统，或由于数据是从备份介质中通过应用系统恢复，再导出为文本文件提供，难以在采集的同时直接对数据进行验证。在审计金融企业时，通常都会遇到这种情况。此时，就必须通过利用电子数据计算有关金额，与现有的纸质资料，如有关财务报表、日常统计表的相应金额比对，如有差异，则应进行分析，找出产生差异的原因。有时候金额有差异，并不表明数据一定不真实，还应进一步分析产生差异的原因。例如，如果获取了某银行明细账户的数据，则可以统计出年末总账余额，将统计出的数据与该行年末业务状况表的相应总账科目余额相比较，如通过电子明细账统计出的总账科目余额小于业务状况表的金额，不能立即得出数据不完整的结论，因为可能存在有部分业务未上机，还是手工处理的情况。

10.4.2 数据真实性的验证方法

1. 验证数据库的创建日期

由审计人员和被审计单位的技术人员一起，由审计人员监督、被审计单位技术人员执行，在被审计单位的 Oracle 数据库管理系统中，在 SQL* PLUS 工具中输入如下命令可以查看当前连接的数据库的创建日期。

```
Select Created From V$ Database;
```

结果如图 10-4 所示。

在得到了被审计单位的表空间后，审计人员可以进一步了解表空间的物理存储位置和所占的空间大小，这可以通过下述语句实现：

```
select  tablespace_name  表空间, file_name  物理文件名,
```



图 10-4　查看数据库的创建日期

```
bytes 字节数 from dba_data_files;
```

结果如图 10-5 所示。



图 10-5　查看表空间的物理文件和大小

如果希望将查询结果保存到文本文件中，可以在 SQL* PLUS 中执行如下命令：

```
spool e:\tablespace.txt
select  tablespace_name  表空间, file_name  物理文件名,
bytes 字节数 from dba_data_files;
spool off
```

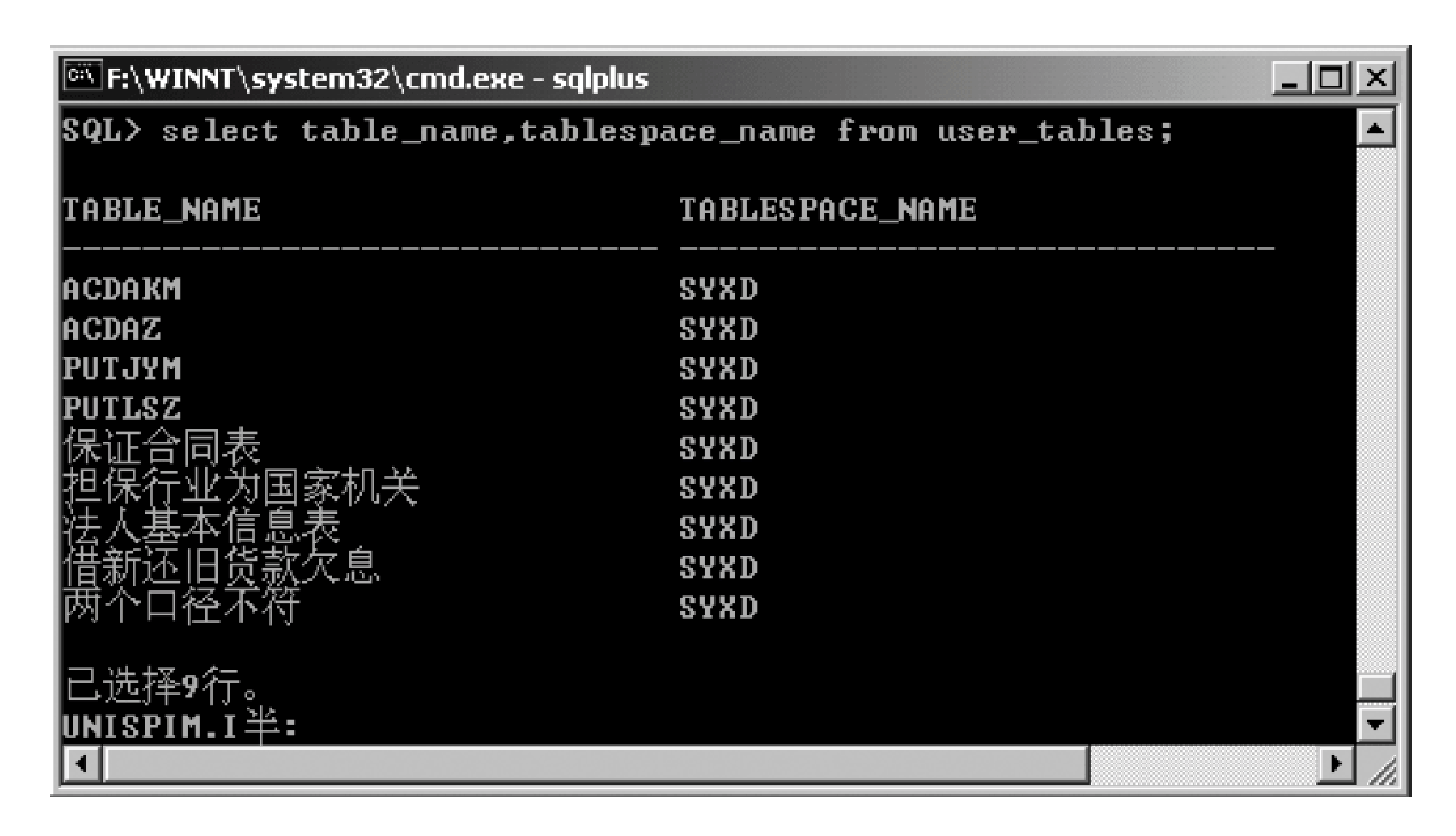
此例是将查询结果以文本方式保存到 e：\ tablespace. txt 文件中，执行完此命令后，审计人员就可以用记事本打开 e：\ tablespace. txt，查看其内容。

如果表空间的大小与审计人员预期的大小相差比较大，则可以对此表空间中的数据的真实性提出质疑。

审计人员还可以通过执行下述语句，得到当前连接用户的全部表和表所在的表空间：

```
SELECT TABLE_NAME, TABLESPACE_NAME FROM USER_TABLES
ORDER BY TABLE_NAME
```

执行结果如图 10-6 所示。


F:\WINNT\system32\cmd.exe - sqlplus
SQL> select table_name,tablespace_name from user_tables;

TABLE_NAME	TABLESPACE_NAME
ACDAKM	SYXD
ACDAZ	SYXD
PUTJYM	SYXD
PUTLSZ	SYXD
保证合同表	SYXD
担保行业为国家机关	SYXD
法人基本信息表	SYXD
借新还旧贷款欠息	SYXD
两个口径不符	SYXD

已选择9行。
UNISPIM.I半:

图 10-6 查看当前连接用户的全部表和所在的表空间

2. 验证两个数据库的结构的差异

在 Oracle 中验证两个数据库的结构的差异可以通过分别比较两个数据库中相似表的结构实现。

示例：验证 PUTJYM 和 PUTJ 表结构的差异，可以使用如下方法。

(1) 首先用下述语句得到 PUTJYM 表的结构，并将结果保存在 d：盘根目录下的 PUTJYM. txt 文件中。

```
sql> spool d:\PUTJYM.txt
sql> desc PUTJYM;
sql> spool off;
```

(2) 然后用下述语句得到 PUTJ 表的结构，并将结果保存在 d：盘根目录下的 PUTJ. txt 文件中。

```
sql> spool c:\PUTJ.txt
sql> desc PUTJ;
sql> spool off;
```

(3) 最后用记事本打开两个文件，对其中的内容进行比较，以查看两个表的结构的差异。

3. 验证数据表是基本表还是视图

审计人员在确定好了需要的数据所在的表之后，在导出表中数据之前，应先在被审计单位的数据库服务器上判断被审计单位提供的数据表是基本表还是视图，并且当数据表是基本表时，审计人员还应该验证这些表的创建日期。

示例：列出某数据库中的全部基本表。

首先连接到要查看的数据库表空间，然后在 SQL* PLUS 中执行下述语句：

```
select table_name from user_tables;
```

结果如图 10-7 所示。

若要将结果保存到一个文本文件中，执行如下语句：

```
sql> spool c:\tab_name.txt
sql> select table_name from user_tables;
sql> spool off;
```

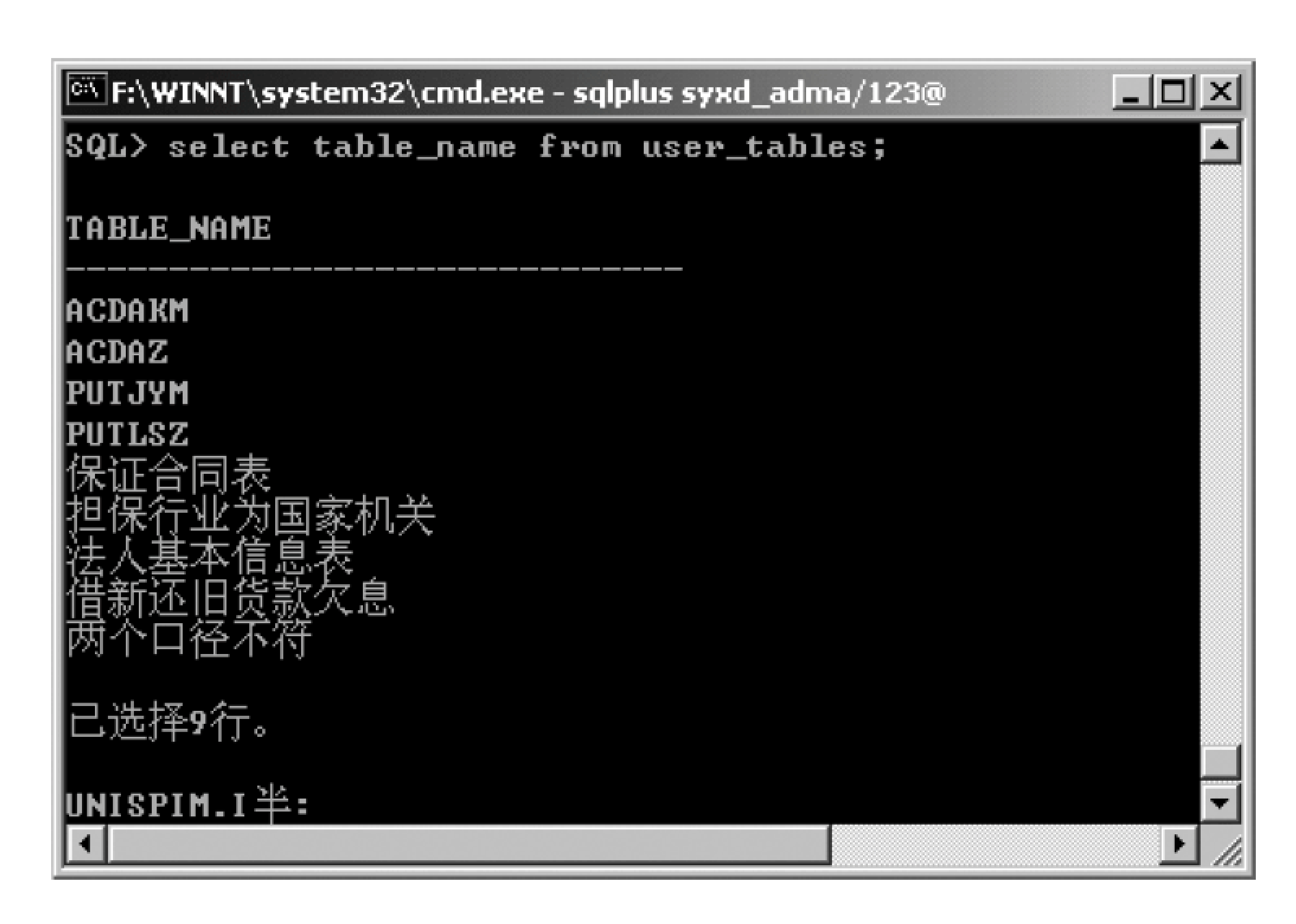


图 10-7　列出数据库中的全部基本表

如果被审计单位提供的表不在这个范围内，则说明他们提供的有可能是视图，审计人员可以进一步执行下述命令验证被审计单位提供的对象是否视图。

示例：列出数据库中的全部视图。

```
SELECT VIEW_NAME FROM USER_VIEWS;
```

执行结果与图 10-8 所示类似。



图 10-8　查看全部视图

如果被审计单位提供的对象是基本表，则应该进一步验证这些基本表的创建日期，以此判断被审计单位提供的数据表是否新近创建的表。

示例：列出数据库中全部基本表和视图的创建日期。

```
Select object_name,object_type,created from user_objects
```

结果与图 10-9 所示类似。

也可以将查询结果保存到文本文件中。

示例：列出数据库中全部基本表和视图的创建日期，并将结果保存到 d：盘根目录下的 info. txt 文件中。

```
spool d:\info.txt
```

```
F:\WINNT\system32\cmd.exe - sqlplus
SQL> select object_name,object_type,created from user_objects;

OBJECT_NAME
--------------------------------------------------------------------------------

OBJECT_TYPE          CREATED
-------------------  ----------
ACDAKM
TABLE                13-5月 -04

ACDAZ
TABLE                13-5月 -04

PUTJYM
TABLE                13-5月 -04


OBJECT_NAME
--------------------------------------------------------------------------------

OBJECT_TYPE          CREATED
-------------------  ----------
PUTLSZ
UNISPIM.I半:
```

图 10-9　列出数据库中全部表和视图的创建日期

```
select object_name,object_type,created from user_objects;
spool off
```

如果被审计单位的数据库中的表或视图比较多，在利用上述语句查找某数据表是否数据库中的基本表或视图时就会比较麻烦，这时我们可以直接对被审计单位提供的数据表对象的类型和创建日期进行验证。

示例：验证“保证合同表”的类型和创建日期。

```
Select object_name,object_type,created from user_objects
Where object_name ='保证合同表';
```

结果如图 10-10 所示。



图 10-10　保证合同表的创建日期

10.4.3　数据完整性的验证

在得到了要审计的数据所在的表之后，审计人员可以进一步查看表中的记录数，将得到的记录数与审计人员估算的记录数进行比较，以验证表数据的完整性。

在 SQL* PLUS 中输入：

```
select count(*) from 表名;
```

可以得到指定表的记录数。

示例：查看“保证合同表”所包含的记录数。

```
select count(*) from 保证合同表;
```

执行结果与图 10-11 所示类似。



图 10-11　得到保证合同表的记录数

参考文献

[1] SOARES S. Big Data Governance: An Emerging Imperative[M]. [S. l.]: MC Press, 2013.

[2] VINCENZO M. Big Data and Analytics—Strategic and Organizational Impacts[M]. Berlin: Springer, 2015.

[3] 刘汝焯. 计算机审计质量控制模型：第 2 版[M]. 北京：清华大学出版社，2016.

[4] BALLARD C. Information Governance Principles and Practices for a Big Data[EB/OL]. http://www. redbooks . ibm. com/redbooks/pdfs/sg248165. pdf.

[5] MCDANIEL P, PAPERNOT N, CELIK Z B. Machine Learning in Adversarial Settings[J]. IEEE Security & Privacy, 2016,14(3): 68-72.

附录 A　Tableau 10.0 简介

A.1　Tableau 工作区

在首次进入 Tableau 或打开 Tableau 但没有指定工作簿时，会显示初始界面，如图 A-1 所示。



图 A-1　Tableau 的初始界面

初始界面的左边是“连接”窗格，在此可以选择连接任何所需要的数据源。右边列出了最近使用的工作簿、已保存的数据连接、示例工作簿等。

Tableau 工作区是制作视图、设计仪表板、生成故事、发布和共享工作簿的工作环境，包括工作表工作区、仪表板工作区和故事工作区，也包括公共菜单栏和工具栏。

- 工作表(work sheet)：又称视图(visualization)，是可视化分析的最基本单元。
- 仪表板(dashboard)：是多个工作簿和一些对象(如图像、文本、网页等)的组合，可以按照一定方式对其进行组织和布局，用于揭示数据关系和内涵。
- 故事(story)：是按顺序排列的工作表或仪表板的集合，故事中各个单独的工作表或仪表板称为“故事点”。可以用故事向用户叙述某些事实，或者以故事方式展示各事实之间的上下文或事件发展关系。

A.1.1　工作表工作区

在 Tableau 连接好数据源之后，即进入工作表工作区，工作表工作区如图 A-2 所示。该工作区包含的主要部件如下。

(1) 数据窗口。数据窗口位于工作表工作区的左侧，图 A-2 中最左边的框框出的部分即为数据窗口。数据窗口包含“数据”和“分析”两个选项卡。

“数据”选项卡中包含以下三部分内容。

- 数据源显示框：包含当期使用的数据源及其他可用的数据源。
- 维度列表框：包含数据源中文本、日期等离散型数据的字段。
- 度量列表框：包含可用于聚合的连续数据的字段。

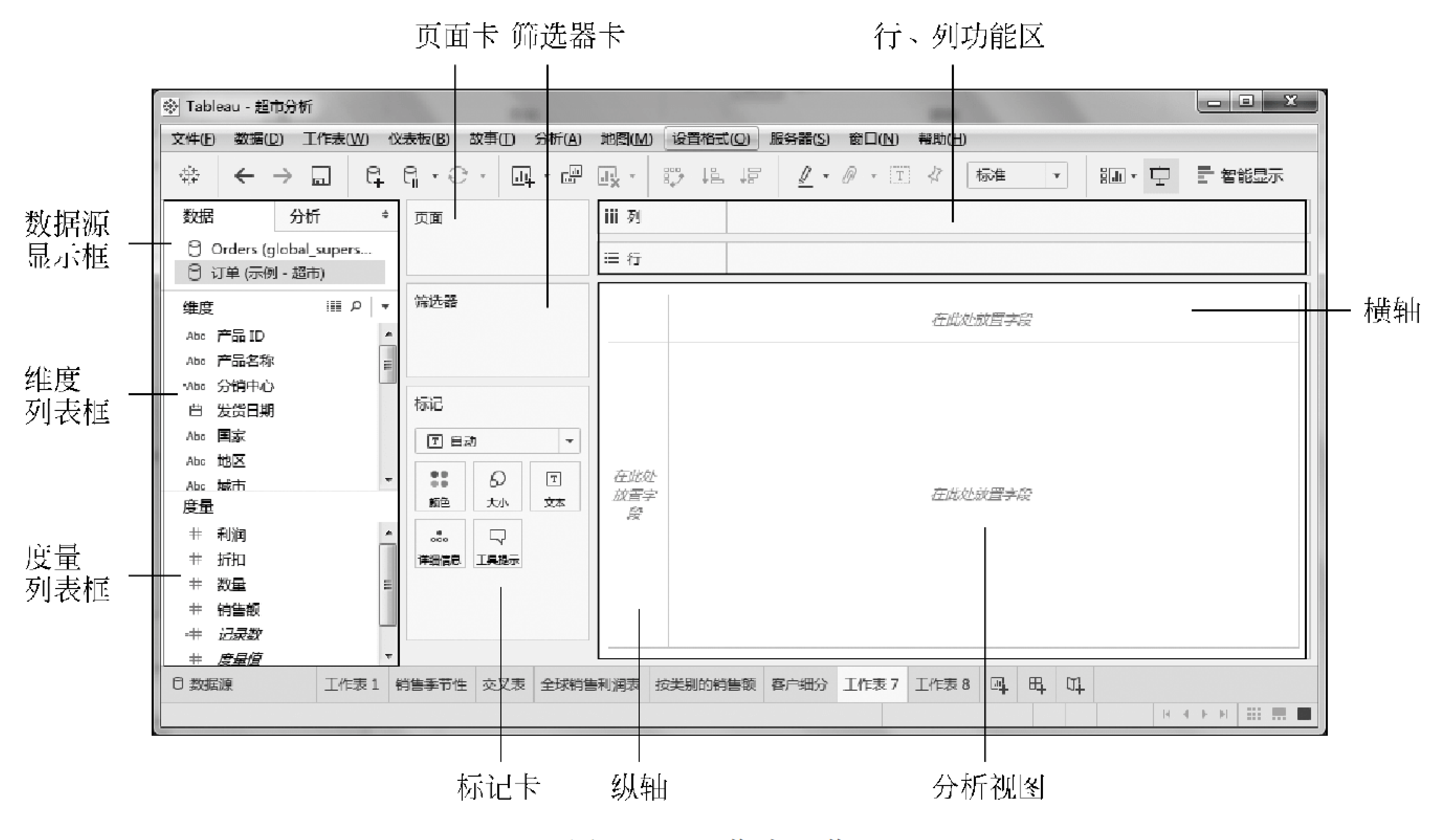


图 A-2　工作表工作区

“分析”选项卡的样式如图 A-3 所示。分析窗格包含菜单中常用的分析功能，便于快速使用。“分析”窗格中主要包含如下几部分内容。

- 汇总：提供常用的参考线(常量线)、平均线、含四分位的中值、盒须图和合计等，可直接将这些拖放到视图中应用。
- 模型：提供常用的分析模型，包括含 95% CI 的平均值、95% CI 的中值、趋势线、预测和群集。
- 自定义：提供参考线、参考区间、分布区间和盒须图的快捷使用。

数据窗口的右边按从上到下有三个卡，分别是页面卡、筛选器卡和标记卡。

(2) 页面卡：可在此功能区中基于某个维度的成员或某个度量的值将一个视图拆分为多个视图。

(3) 筛选器卡：用于指定要包含和排除的数据，所有经过筛选的字段都显示在筛选器卡上。

（4）标记卡：控制视图中的标记的属性，包括一个标记类型选择器，可以在其中指定标记类型，如条形图、线、圆等，此外还包含颜色、大小、标签、文本、详细信息、工具提示等。

（5）行、列功能区：行功能区用于创建行，列功能区用于创建列，可以将任意数量的字段放置在这两个功能区中。

（6）智能显示：智能显示包含的内容如图 A-4 所示。通过智能显示，可以基于视图中已经使用的字段以及数据窗口中选择的任意字段来创建视图。Tableau 会自动评估选定的字段，然后在智能显示中突出显示与数据最相符的可视化图表类型。

（7）工作表视图区：创建和显示视图的区域，一个视图就是行和列的集合，包括的组件有标题、轴、区、单元格和标记。此外，还可以选择显示标题、说明、字段标签、摘要和图例等。

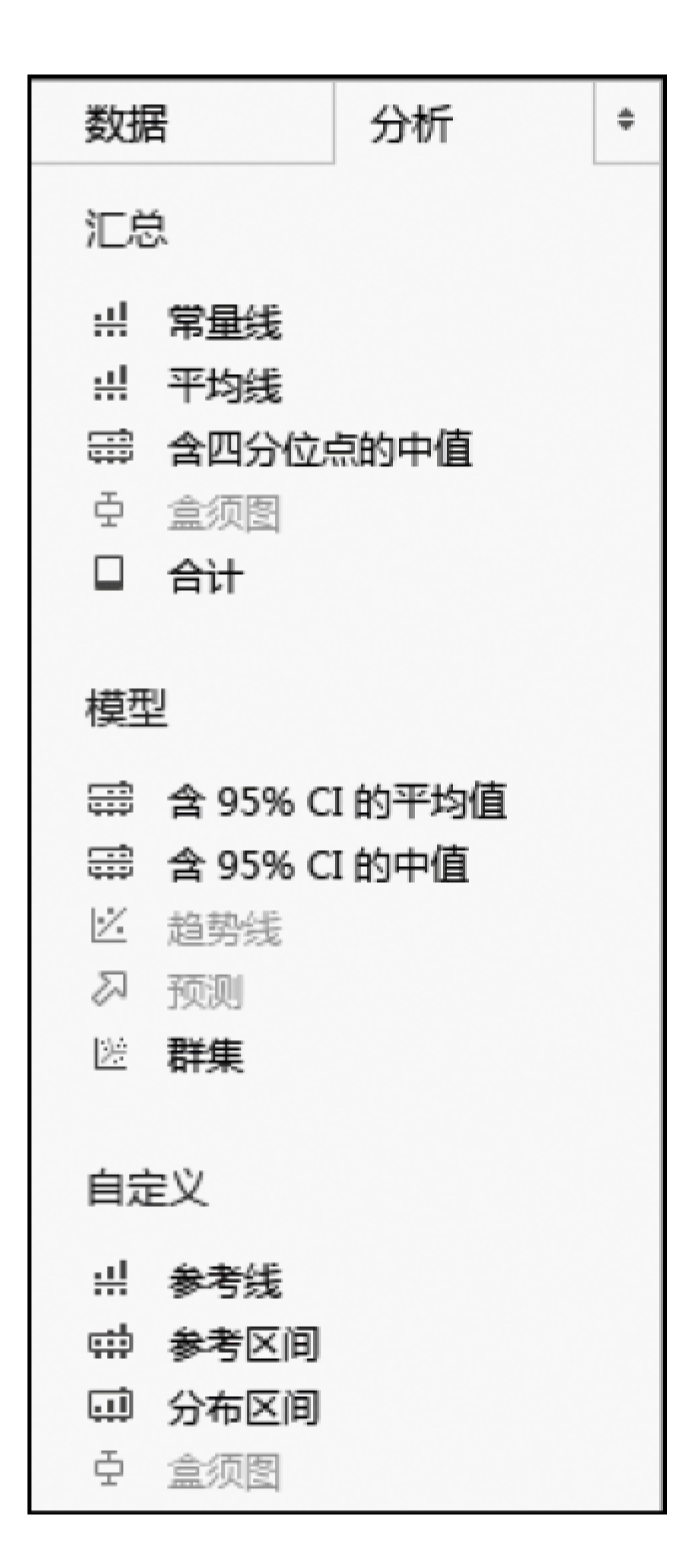


图 A-3 “分析”选项卡



图 A-4 “智能显示”选项卡

A.1.2 仪表板工作区

仪表板工作区使用布局容器把工作表和诸如图片、文本、网页类型的一些对象按一定的布局方式组织在一起。在工作区页面单击新建仪表板图标，或者选择“仪表板”菜单下的“新建仪表板”，即可打开仪表板工作区，如图 A-5 所示。

仪表板包含的主要部件如下（按从左到右，从上到下的顺序介绍）。

（1）仪表板窗格：列出了当期工作簿中创建的所有工作簿，可以选中工作表并将其从仪表板窗格拖放到右侧的仪表板区域中。仪表板区域中的灰色区域将指示出该工作表可放置的位置。

（2）仪表板对象窗格：包含仪表板支持的对象，如文本、图形、网页和空白区域。从

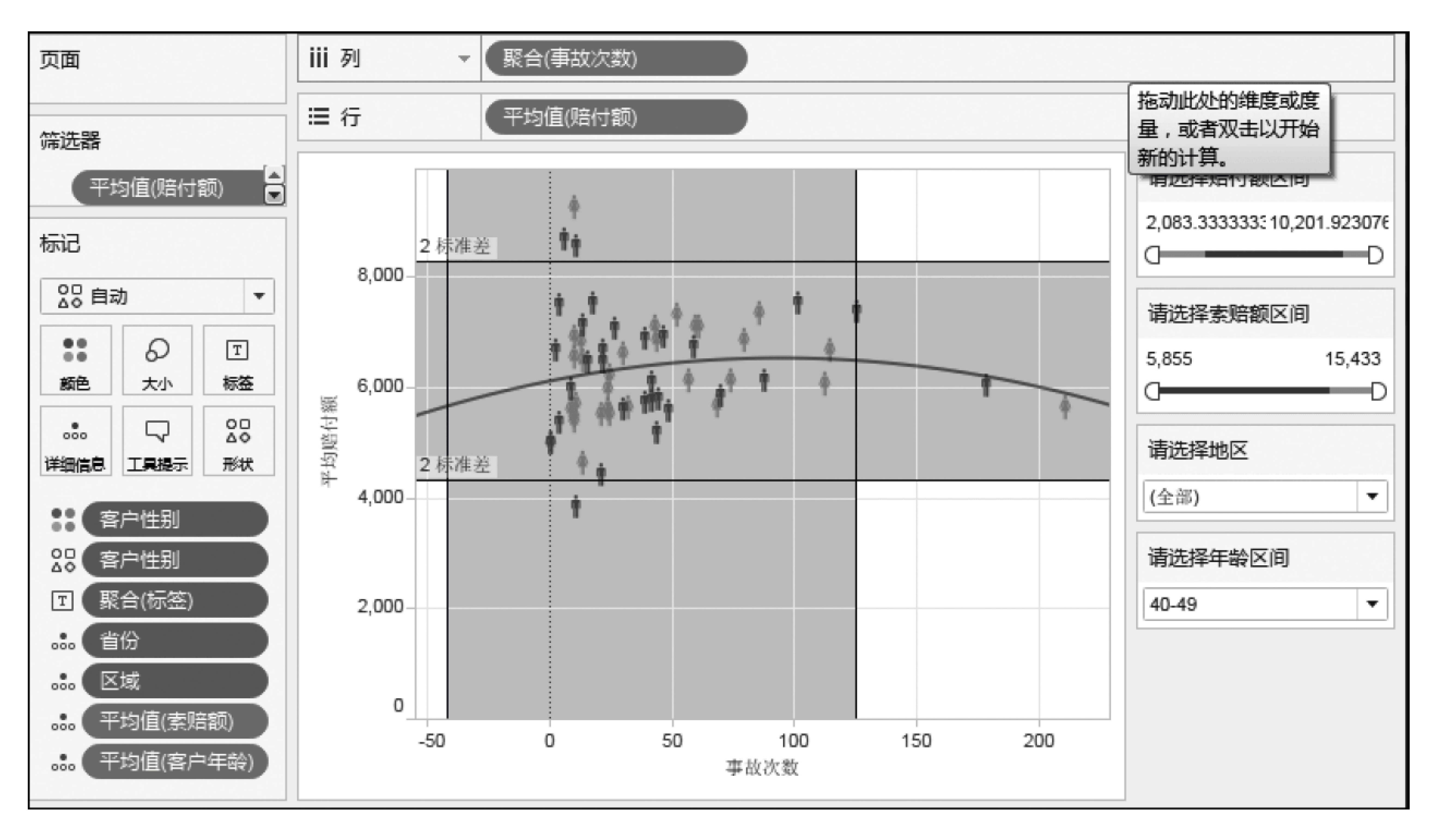


图 A-5　仪表板工作区

仪表板窗格拖放所需对象至仪表板窗格中，可以添加仪表板对象。

（3）平铺和浮动：决定了工作表和对象被拖放到仪表板后的效果和布局方式。默认情况下，仪表板使用平铺布局，这意味着每个工作表和对象都排列在一个分层网格中。可以将布局改为“浮动”以允许视图和对象重叠。

（4）布局窗格：以树形结构显示当前仪表板中用到的所有工作表集对象的布局方式。

（5）仪表板设置窗格：设置创建的仪表板的大小，也可以设置是否显示仪表板标题。仪表板的大小可以从预定义的大小中选择一个，或以像素为单位设置自定义大小。

（6）仪表板视图区：是创建和调整仪表板的工作区域，可以添加工作及各类对象。

A.1.3　故事工作区

故事是 Tableau 8.2 之后新增加的图形，一般将故事用在演示工具，按顺序排列视图或仪表板。在工作区页面单击新建故事图标，或者选择“故事”菜单下的“新建故事”，即可打开故事工作区，如图 A-6 所示。

故事工作区包含的主要部件如下（按从左到右，从上到下的顺序介绍）。

（1）仪表板和工作表窗格：列出在当期工作簿中创建的工作表和仪表板，将其中一个工作表或仪表板拖放到故事区域，即可创建故事点。

（2）说明：说明是可以添加到故事点中的一种特殊类型的注释。若要连接说明，只需双击此处。可以向一个故事点添加多个说明，可以将说明放置到故事中的任何合适的位置。

（3）导航器设置：设置是否显示导航框中的后退/前进按钮。

（4）故事设置窗格：设置创建的故事的大小，也可以设置是否显示故事标题。故事

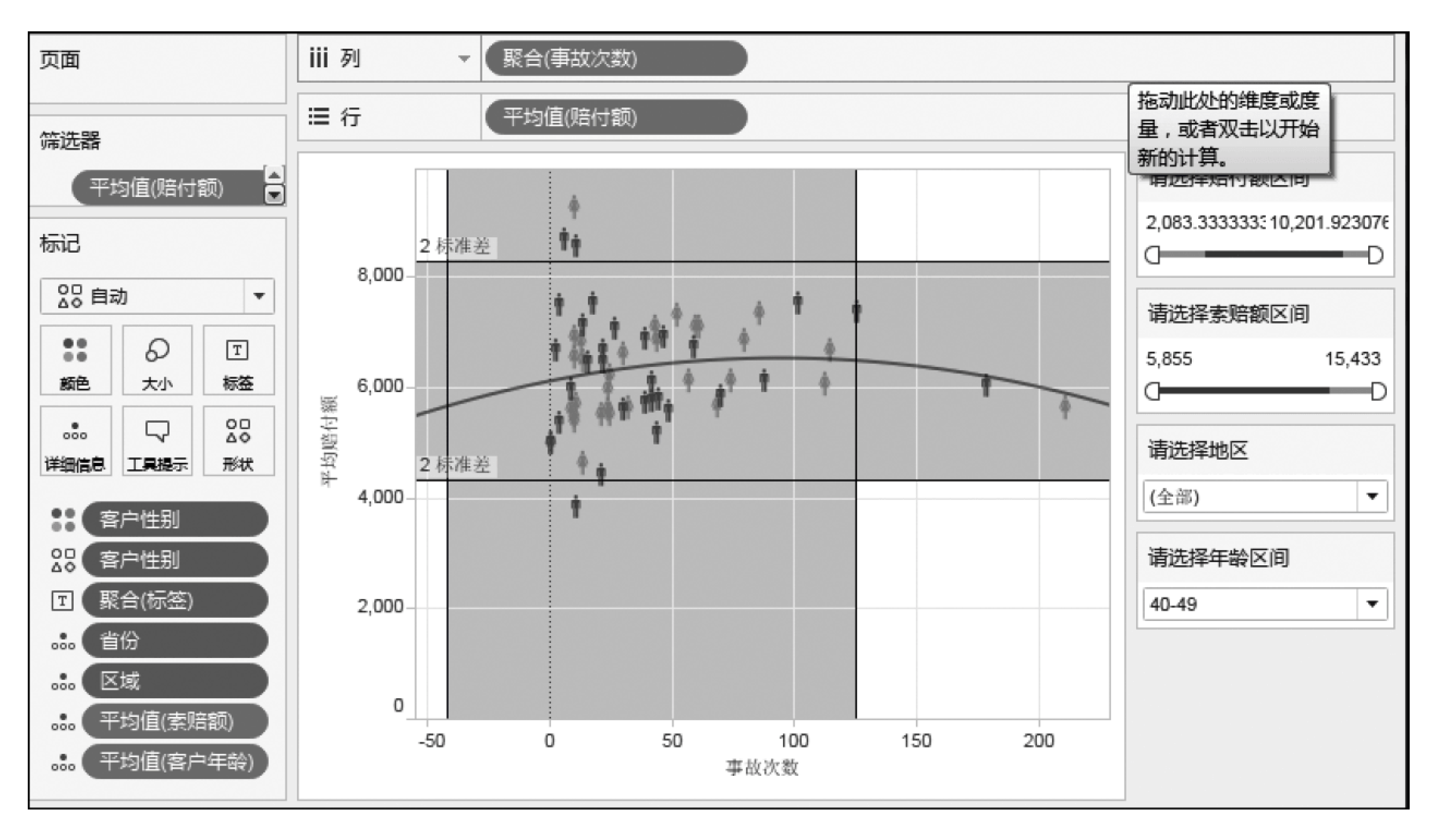


图 A-5　仪表板工作区

仪表板窗格拖放所需对象至仪表板窗格中,可以添加仪表板对象。

(3) 平铺和浮动:决定了工作表和对象被拖放到仪表板后的效果和布局方式。默认情况下,仪表板使用平铺布局,这意味着每个工作表和对象都排列在一个分层网格中。可以将布局改为“浮动”以允许视图和对象重叠。

(4) 布局窗格:以树形结构显示当前仪表板中用到的所有工作表集对象的布局方式。

(5) 仪表板设置窗格:设置创建的仪表板的大小,也可以设置是否显示仪表板标题。仪表板的大小可以从预定义的大小中选择一个,或以像素为单位设置自定义大小。

(6) 仪表板视图区:是创建和调整仪表板的工作区域,可以添加工作及各类对象。

A.1.3　故事工作区

故事是 Tableau 8.2 之后新增加的图形,一般将故事用在演示工具,按顺序排列视图或仪表板。在工作区页面单击新建故事图标,或者选择“故事”菜单下的“新建故事”,即可打开故事工作区,如图 A-6 所示。

故事工作区包含的主要部件如下(按从左到右,从上到下的顺序介绍)。

(1) 仪表板和工作表窗格:列出在当期工作簿中创建的工作表和仪表板,将其中一个工作表或仪表板拖放到故事区域,即可创建故事点。

(2) 说明:说明是可以添加到故事点中的一种特殊类型的注释。若要连接说明,只需双击此处。可以向一个故事点添加多个说明,可以将说明放置到故事中的任何合适的位置。

(3) 导航器设置:设置是否显示导航框中的后退/前进按钮。

(4) 故事设置窗格:设置创建的故事的大小,也可以设置是否显示故事标题。故事



图 A-6 故事工作区

的大小可以从预定义的大小中选择一个,也可以以像素为单位设置自定义大小。

(5) 导航框:用户进行故事点导航的窗格,可以利用左侧或右侧的按钮顺序切换故事点,也可以直接单击故事点进行切换。

(6) 新空白点按钮:单击此按钮可以创建新故事点,使其与原来的故事点有所不同。

(7) 复制按钮:可以将当前故事点用作新故事点的起点。

(8) 说明框:通过说明为故事点或者故事点中的工作表或仪表板添加注释的文本框。

(9) 故事视图区:创建故事的工作区域,可以添加工作表、仪表板或说明框对象。

A.2 Tableau 的文件管理

Tableau 文件有多种类型,如工作簿、打包工作簿、数据提取、数据源和书签等,用于保存工作成果和数据源。表 A-1 列出了 Tableau 的文件类型。

- 工作簿文件(.twb):占用空间小,默认的保存方式,包含所有工作表及连接信息,但不包含数据。

- 打包工作簿文件(.twbx):占用空间可能非常大,是一个 zip 文件,包含所有工作表、连接信息以及全部本地资源(如本地数据源、背景图片、自定义地理编码等)。这种格式最适合对工作进行打包以便与不能访问该数据的其他人共享。
- 数据源文件(.tds):占用空间极小。数据源文件是快速连接经常使用的数据源的快捷方式。数据源文件不包含实际数据,只包含新建数据源所必需的信息以及在数据窗口中所做的修改,如默认属性、计算字段、组、集等。
- 数据源文件(.tdsx):占用空间小。如果连接的数据源不是本地数据源,则.tdsx 文件的内容与.tds 文件相同;如果连接的数据源是本地数据源,则.tdsx 文件不但包含.tds 文件中的所有信息,还包含本地文件数据源(Excel、Access、文本和数据提取)。
- 书签文件(.tbm):通常占用空间比较小。书签包含单个工作表,是快速分享所做工作的一种简便方式。
- 数据提取文件(.tde):占用空间可能非常大。数据提取文件是部分或整个数据源的一个本地副本,可用于共享数据、脱机工作和提高数据库性能。

附录 B　RapidMiner 使用方法简介

RapidMiner 是数据挖掘、机器学习和商业预测分析领域的一款备受用户青睐的软件，用户可以从 www.rapidminer.com 免费下载使用。RapidMiner 与其他数据分析软件相比，具有用户入门快，操作简单的特点，用户使用它不需要任何编程知识，只需通过鼠标拖放，就能完成数据挖掘和分析的功能。

B.1　RapidMiner 的主界面

图 B-1 是 RapidMiner 的主界面，主要包括四个区域。

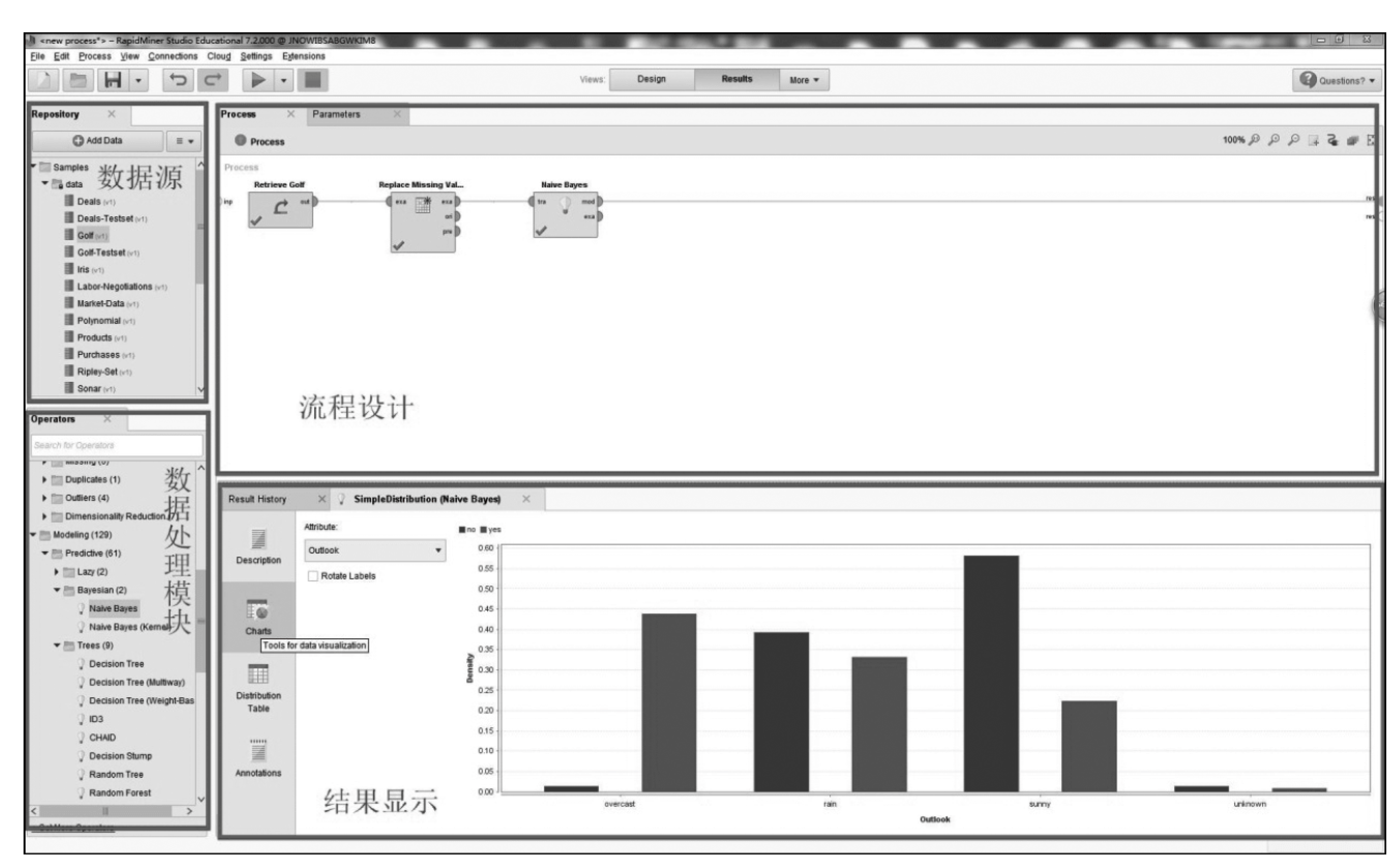


图 B-1　RapidMiner 主界面

1. 数据源区域：数据源区域列出了 RapidMiner 当前可以使用的数据源。RapidMiner 支持多种数据源，包括 Excel、XML、SQL Server、Oracle、MySQL 和 Hadoop 等。如果数据源不存在，可以单击数据源区域上方的“Add Data”按钮来添加。

2. 数据处理模块区域：数据处理模块区域包含各种数据处理和分析方法，它们在 RapidMiner 中又称为“算子”(operator)。RapidMiner 提供了很多数据处理和分析方法，如数据清洗、数据统计、数据分类、聚类分析、关联分析和数据预测等。

3. 流程设计区域：“流程”是 RapidMiner 数据分析的核心内容，数据分析的任务其实就是设计各种流程。流程可以理解成 RapidMiner 中数据分析的过程，数据源以及按照

先后顺序使用的一系列数据分析处理方法就构成了一个流程。图 B-2 所示是一个流程，流程中的数据源和各种处理方法都用一个方块代表，流程中各个方块从左到右的连线代表了数据分析处理的步骤。图 B-2 中的流程从左边的数据源模块开始，从左至右，经过两个数据分析处理模块(分别是“替换缺失的数据”和“决策树”)后，通过右边的“结果输出”端口给出分析的结果。流程中每一个方块的左边和右边分别有一些小的凸起，代表该处理模块的数据输入端口和数据输出端口，左边的凸起是数据输入端口，右边的凸起是数据输出端口。从最左边的数据源模块开始，按照从左至右的顺序，依次把上一个模块的数据输出端口和下一个模块的数据输入端口相连接，直至把最后一个模块的输出端口和流程设计窗口的输出结果端口相连接，就完成了一次流程的设计。

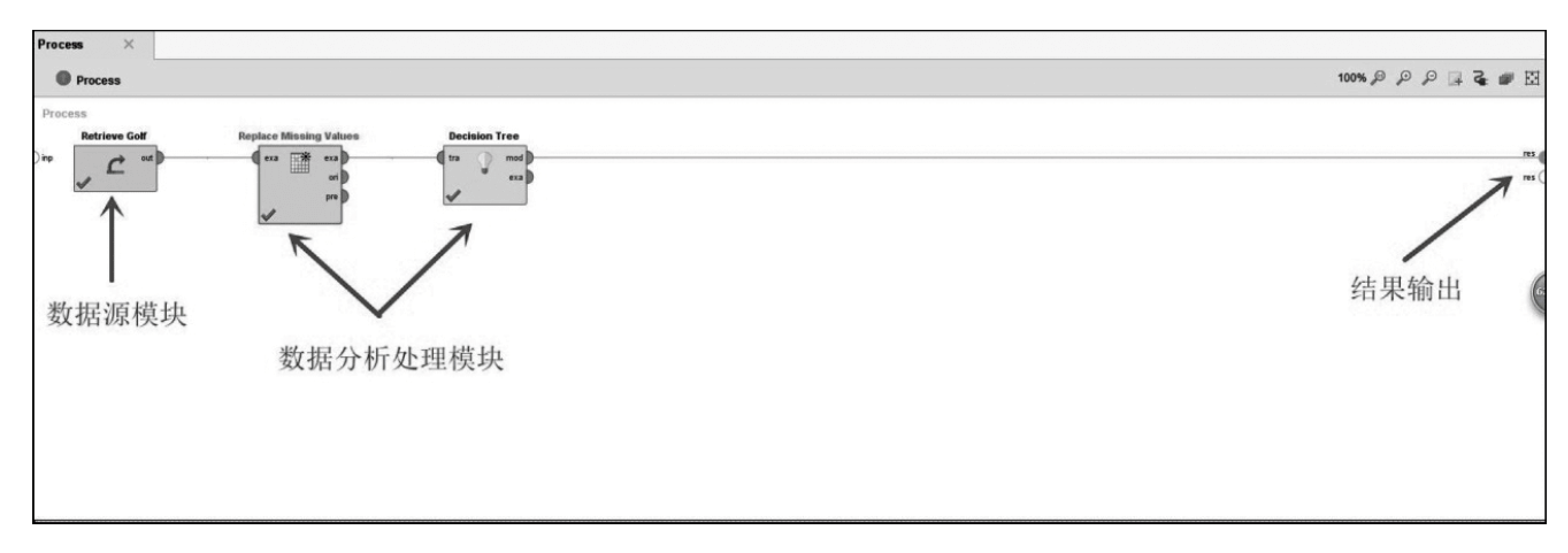


图 B-2　流程

4. 结果显示区域：数据分析的结果在该区域显示。结果显示的形式也非常丰富，包括数据结果和各种统计图表等。

B.2　使用 RapidMiner 分析数据的方法

使用 RapidMiner 分析数据非常简单，不用编写代码，只需像使用积木一样，把数据源和现成的数据分析模块用鼠标拖放到流程设计区域中，前后相连组成一个流程就能完成分析任务。使用 RapidMiner 分析数据的方法可以简单概括成“加载数据，设计流程，配置参数，运行流程”四个步骤。下面以泰坦尼克号乘客数据的决策树分析为例，来简要说明 RapidMiner 分析数据的方法

1. 加载数据

当需要分析新的数据时，可以单击 RapidMiner 主界面中“数据源”区域上方的“Add Data”按钮来添加新的数据源。单击“Add Data”按钮后，出现数据选择对话框，如图 B-3 所示。可以根据数据所在的位置选择在本机还是其他数据库。

2. 设计流程

当加载数据完成后，开始设计流程对数据进行分析。如图 B-4 所示，我们首先用鼠标从“数据源”中选择“Titanic”数据，然后拖放到流程设计区域，再用鼠标从“数据分析处理模块”中选择并拖放如下三个模块到流程设计区域：替换缺失的数据值、设置标签和决策树分析。按照数据处理的步骤把这四个模块从左到右依次连接起来，把最后一个“决策树分析”模块的数据输出端口和最右边的“结果输出”端口相连接。

先后顺序使用的一系列数据分析处理方法就构成了一个流程。图 B-2 所示是一个流程，流程中的数据源和各种处理方法都用一个方块代表，流程中各个方块从左到右的连线代表了数据分析处理的步骤。图 B-2 中的流程从左边的数据源模块开始，从左至右，经过两个数据分析处理模块(分别是“替换缺失的数据”和“决策树”)后，通过右边的“结果输出”端口给出分析的结果。流程中每一个方块的左边和右边分别有一些小的凸起，代表该处理模块的数据输入端口和数据输出端口，左边的凸起是数据输入端口，右边的凸起是数据输出端口。从最左边的数据源模块开始，按照从左至右的顺序，依次把上一个模块的数据输出端口和下一个模块的数据输入端口相连接，直至把最后一个模块的输出端口和流程设计窗口的输出结果端口相连接，就完成了一次流程的设计。

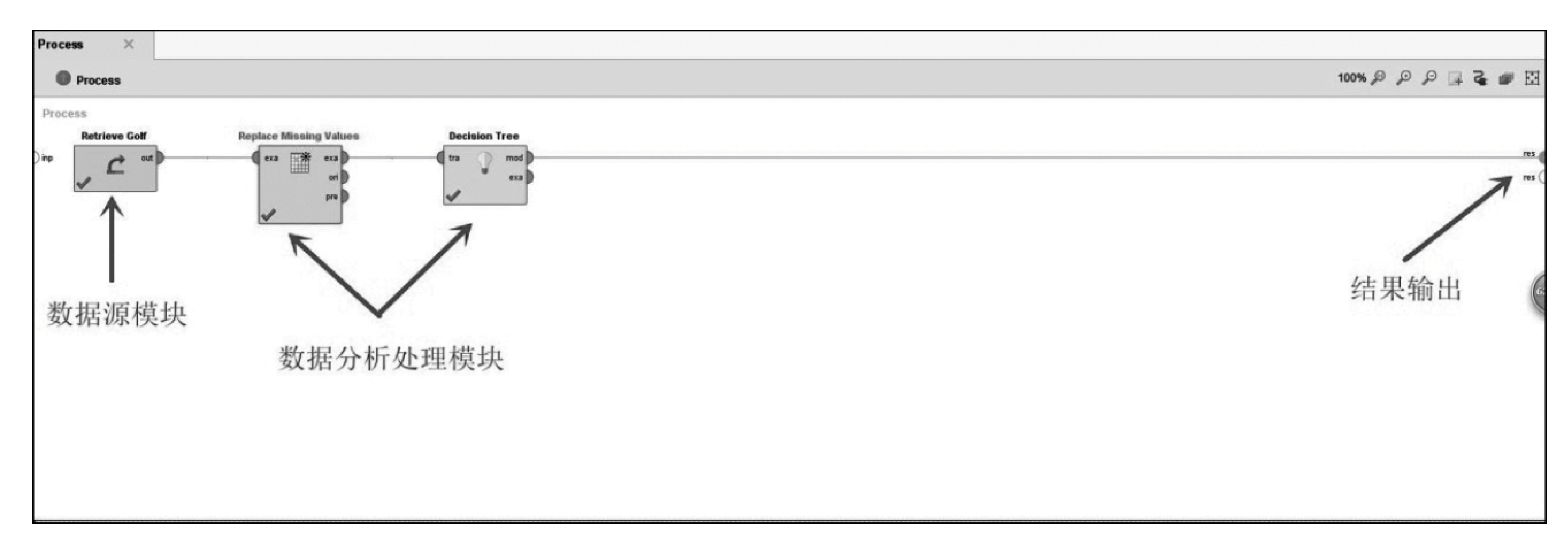


图 B-2　流程

4. 结果显示区域：数据分析的结果在该区域显示。结果显示的形式也非常丰富，包括数据结果和各种统计图表等。

B.2　使用 RapidMiner 分析数据的方法

使用 RapidMiner 分析数据非常简单，不用编写代码，只需像使用积木一样，把数据源和现成的数据分析模块用鼠标拖放到流程设计区域中，前后相连组成一个流程就能完成分析任务。使用 RapidMiner 分析数据的方法可以简单概括成“加载数据，设计流程，配置参数，运行流程”四个步骤。下面以泰坦尼克号乘客数据的决策树分析为例，来简要说明 RapidMiner 分析数据的方法

1. 加载数据

当需要分析新的数据时，可以单击 RapidMiner 主界面中“数据源”区域上方的“Add Data”按钮来添加新的数据源。单击“Add Data”按钮后，出现数据选择对话框，如图 B-3 所示。可以根据数据所在的位置选择在本机还是其他数据库。

2. 设计流程

当加载数据完成后，开始设计流程对数据进行分析。如图 B-4 所示，我们首先用鼠标从“数据源”中选择“Titanic”数据，然后拖放到流程设计区域，再用鼠标从“数据分析处理模块”中选择并拖放如下三个模块到流程设计区域：替换缺失的数据值、设置标签和决策树分析。按照数据处理的步骤把这四个模块从左到右依次连接起来，把最后一个“决策树分析”模块的数据输出端口和最右边的“结果输出”端口相连接。

图 B-3　加载数据

图 B-4　流程设计

3. 配置参数

在设计流程的过程中，我们还可以对每一个处理模块设置不同的参数。只要选择该模块，然后单击“参数”(Parameters)标签，即可设置该模块的相关参数。例如，我们选择“替换缺失的数据值”模块，再单击“参数”(Parameters)标签，即可切换到对该模块的参数设置界面，如图 B-5 所示。



图 B-5　参数配置

4. 运行流程

当设计完流程，配置好各种模块的处理参数后，即可单击菜单中的“运行流程”按钮启动数据分析。“运行流程”按钮如图 B-6 中黑框内所示。

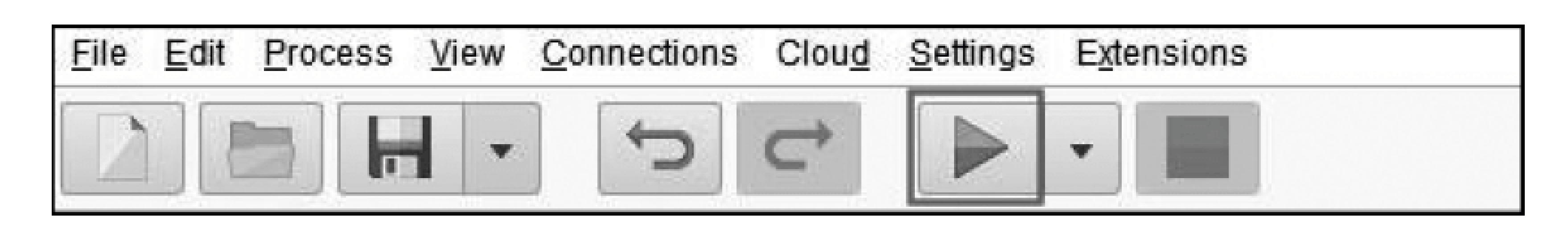


图 B-6　运行流程的按钮

如果流程没有错误，RapidMiner 就会按照从左到右的顺序依次执行流程中的各个处理模块，最后在结果框中显示本流程对数据的分析结果，如图 B-7 所示。

图 B-7　流程对数据的分析结果

从上述步骤可以看到，用 RapidMiner 分析数据不用编写代码，只需像使用积木一样，把数据源和现成的数据分析模块前后组成一个流程就能完成分析任务。